

Proefschrift voorgelegd tot het bekomen van de graad van Doctor in de Moraalwetenschappen
Promotor: Prof. dr. Johan Braeckman

Promotor Prof. dr. Johan Braeckman
 Vakgroep Wijsbegeerte en Moraalwetenschap

Decaan Prof. dr. Marc Boone
Rector Prof. dr. Anne De Paepe

Nederlandse vertaling:

Vrij en gelijk geboren? Over ethische consistentie en dierenrechten



Faculteit Letteren & Wijsbegeerte

Stijn Bruers

Born free and equal?

On the ethical consistency of animal equality

Proefschrift voorgelegd tot het behalen van de graad van
Doctor in de moraalwetenschappen

2014

Acknowledgements

First of all, I would like to thank my supervisor Prof. Dr. Johan Braeckman. He allowed me to explore many new paths in ethics and philosophy and he took time to guide me through the research. His assistance and advice were precious.

Second, I owe gratitude to Prof. Dr. Tom Claes and Tim De Smet for comments and Dianne Scatrine and Scott Bell for proofreading. Thanks to Gitte for helping me with the lay-out and cover. Thanks to the University of Ghent for giving opportunities, knowledge and assistance.

From all philosophers I know, perhaps Floris van den Berg has ethical ideas closest to mine. I enjoyed our collaboration and meetings with him. Also the many discussions with animal rights activists of Bite Back, with the participants at the International Animal Rights Conferences and Gatherings and with the many meat eaters I encountered during the years allowed me to refine my theories. Furthermore, I am grateful to anonymous reviewers for some useful comments on my research papers. Let me also thank the many authors listed in the bibliography, because without their work my hands would be empty.

And finally, of course, a warm thank you to all my family, friends and fellow activists for their support and friendship: my parents, my sister (and her little Floris), An, An-Katrien, Rosie, Dennis, Tim, Benjamin, Ralph, Tobias,...

Abstract

This dissertation investigates the possibility of constructing a consistent ethical system that offers clear notions of equality and incorporates an animal ethic. The first part is more meta-ethical in nature, reflecting on notions such as moral intuitions, universalism, consistency and coherence. It demonstrates that moral illusions might exist and offers a method to discover such moral illusions.

The second part turns to normative ethics, dealing with principles of welfare, justice and basic rights. It tackles problems ranging from population ethics to non-ideal theory.

Finally, the third part moves to applied (animal) ethics. In analogy to optical illusions, I demonstrate that speciesism is not only a kind of prejudicial discrimination but also a moral illusion: an obstinate intuitive judgment that is inconsistent with a coherent system. The third part also tackles the predation problem and the sentience problem in animal ethics.

The end result of this work is a pluralist principlist ethical system that can be captured in a metaphor of five moral fingers working together as *the moral hand*. This moral hand is a constructed, coherent ethical system of five universalized ethical principles based on strong moral intuitions. The *thumb* represents the principle of universalism, which is a basic ingredient of coherentism, and generates an anti-discrimination rule. The *index finger* symbolizes a consequentialist welfare ethic, based on the coherence of impartiality and empathy. The *middle finger* is the mere means principle of a deontological rights ethic: humans (and animals) have a right not to be used as merely means to someone else's ends. This principle captures a lot of moral intuitions that pop up in famous dilemmas. A fourth principle, the *ring finger*, refers to the value of biodiversity and adopts some elements of carnism, the opposite of veganism as ideology. This fourth principle solves the predation problem and is coherent with some other moral intuitions. Finally, the *little finger* represents the principle of tolerated partiality which can be used in some final moral dilemmas. With these five fingers of ethics, we can grasp the moral problem of consuming animal products, and answer the question whether veganism is a moral duty.

Table of Contents

Introduction	1
Summary: towards a coherent theory of animal equality	9
 <i>Part 1 Ethical consistency.....</i>	<i>23</i>
Chapter 1 The basic elements.....	25
1.1 The input data: moral intuitions	25
1.2 The method: rule universalism.....	30
1.2.1 Universalizations made by meat eaters	35
1.2.2 Universalizations made by animal ethicists.....	38
Chapter 2 The goal: consistency and coherence	41
Chapter 3 The problem: moral illusions.....	47
3.1 Optical illusions.....	48
3.2 Moral illusions.....	54
3.3 An example of moral illusions in the trolley dilemma	57
3.4 Is the deontological right a moral illusion?	63
3.5 Heuristics in thought experiments	65
Summary of part one.....	69
 <i>Part 2 Theories of equality</i>	<i>73</i>
Chapter 4 Impartiality and prioritarian equality.....	77
4.1 Contractarianism universalized	77
4.2 From feelings and well-being to the value of life	79
4.2.1 Affective qualia: from experienced feelings to experienced pleasure	79
4.2.2 The importance of preferences: from experienced pleasure to momentaneous well-being.....	80

4.2.3	The problem of interpersonal comparability: from individual well-being to comparable momentaneous well-being	82
4.2.4	The lifetime perspective: from momentaneous well-being to the value of life	85
4.2.5	Personal identity and psychological continuity	89
4.3	The maximin principle	92
4.4	The quasi-maximin principle and prioritarianism.....	94
4.5	Applications of the quasi-maximin theory.....	96
4.5.1	Rawls' theory of justice	96
4.5.2	Responsibility and desert.....	97
4.5.3	Future orientation and restorative justice	101
4.6	Aversions behind the veil of ignorance: a mathematical description for a theory of justice	102
4.6.1	Why a mathematical model?	102
4.6.2	The mathematics of consequentialist welfare ethics	103
4.6.3	The impartial observer behind the veil of ignorance	106
4.6.4	Deriving the welfare function behind the veil of ignorance	108
4.6.5	The reflection effect and risk neutrality for negative well-being levels.....	114
4.6.6	Loss aversion.....	114
4.6.7	Problematic properties of number-dampened prioritarianism.....	121
4.6.8	Intermezzo: a more complex formulation to solve the replaceability problem.....	127
4.6.9	Uncertainty aversion	134
4.6.10	Prioritarian theories for lotteries	140
4.6.11	Democratic impartial preferences of moral agents	144
4.7	Summary	147
Chapter 5	Partiality and tolerated choice equality.....	149
5.1	Tolerated choice equality.....	150
5.2	To whom applies the tolerated choice equality?	151
5.3	Tolerated choice equality and equality of opportunity	153
Chapter 6	Basic right equality	155
6.1	Moral dilemmas and strong moral intuitions	155
6.1.1	A first approach: uncertainty aversion.....	156
6.1.2	Tentative ethical principles.....	157
6.2	The basic right and the mere means principle	159
6.3	When is the basic right violated?	163
6.3.1	Two words, two conditions.....	163
6.3.2	Conclusion	171
6.4	Who gets the basic right?	172
6.5	How strong is the basic right?	177
6.5.1	Intermezzo: combining the basic right with the prioritarian theory	179
6.6	The extended mere means principle	181
6.6.1	Doing versus allowing.....	182

6.6.2	Tolerated partiality and imperfect duties	183
6.6.3	The asymmetry of procreational duties	185
6.7	Application: the least harm principle and vegetarianism.....	186
Chapter 7	Summary: principles of equality and further refinements	189
7.1	Equality and veganism	191
7.2	Ideal and non-ideal theory: applying the universalist imperative	193
7.2.1	The argument of futility	196
7.2.2	Tit-for-what?	196
7.2.3	Prohibition laws.....	199
7.2.4	Self-defense against culpable attackers and innocent threats.....	200
7.2.5	Summary.....	203
7.3	Formal equality, discrimination and hierarchic dualism	204
Part 3	<i>Animal ethics</i>	207
Chapter 8	Speciesism as a moral illusion	209
8.1	The current situation: patho-anthropocentrism.....	209
8.2	Moral illusions and discrimination	210
8.3	How do we know whether speciesism is a moral illusion?	212
8.4	Five arguments against the species boundary	216
8.5	Five arguments in favor of sentience	219
8.6	Speciesism and cognitive impenetrability.....	223
8.7	Psychological background theories: human prejudices and essentialism.....	225
8.8	Speciesism as a moral heuristic.....	228
8.8.1	The heuristics hypothesis	229
8.8.2	Time and knowledge constraints.....	231
8.8.3	Fear of a slippery slope	232
8.8.4	The emotional cost of excluding atypical humans.....	234
8.8.5	The importance of sentience	236
8.9	Summary	240
Chapter 9	The sentience problem	243
9.1	The scientific problem	243
9.2	The ethical problem	246
Chapter 10	The predation problem.....	249
10.1	Invalid solutions to the prey problem.....	251
10.2	A hypothetical solution to the prey problem.....	252
10.3	Invalid solutions to the difference problem.....	253
10.4	A first hypothetical solution to the difference problem: the 3-N-principle.....	259
10.5	The value of biodiversity	261
10.6	Some further tests for the 3-N principle.....	266
10.7	A second hypothetical solution to the difference problem: behavioral fairness	270

10.8 Summary	273
Chapter 11 The property problem and the harvest problem	275
11.1 Habitat destruction	276
11.2 Animals killed in harvest	278
 Part 4 Epilogue.....	 281
Chapter 12 The moral hands	283
12.1 The moral hand of normative ethics: five principles of a complete and coherent ethic	283
12.1.1 Five principles of equality	285
12.1.2 Applications of the five fingers	286
12.1.3 Intermezzo: maps of the moral landscape	291
12.2 A second moral hand of meta-ethics	295
12.3 The impossible triangle of the meat eater	299
12.4 Argumentation scheme for veganism	302
Where to go from here? Questions for future research.....	305
Nederlandstalige samenvatting: dieren, dilemma's en discriminatie	309
Inleiding	309
1. De duim: het universalismeprincipe, coherentie en morele illusies	311
1.1. Coherentie als universalisme plus consistentie	312
1.2. Morele illusies	313
1.3. De universalistische duimregel en antidiscriminatie	314
1.4. Speciesisme als morele illusie.....	315
2. De wijsvinger: het welzijnsprincipe, onpartijdigheid en empathie.....	320
2.1. Onpartijdigheid	320
2.2. Empathie	322
3. De middelvinger: het middelprincipe, respect en waardigheid	323
4. De ringvinger: het biodiversiteitsprincipe en het predatiedilemma	326
4.1. Het predatieprobleem.....	326
4.2. Het 3-N-principe	327
4.3. De waarde van biodiversiteit	328
4.4. Twee andere oplossingen voor het predatieprobleem	330
5. De pink: het partijdigheidsprincipe en getolereerde keuzegelijkwaardigheid	331
Conclusie: veganisme als morele plicht?	333
Appendix: a review and systematization of the trolley problem	337
Abstract	337
Introduction	337
The trolley dilemmas	340
Six algorithmic accounts	342
Group A: the 'mere means' accounts	343

Group B: the ‘same threat’ accounts	344
Group C: the ‘causal chain’ accounts	346
Seven psychological accounts	349
Four invalid accounts.....	354
Conclusion and further research	355
Bibliography	357

Introduction

The discussion about the moral status of non-human animals (hereafter: animals), and the use of animals for food, clothing, entertainment or research, has a long history that goes back to Ancient Greece (Pythagoras, Plutarch). From time to time the problem resurges throughout the centuries (Leonardo Da Vinci; Jeremy Bentham; Oswald, 1791; Ritson, 1802; Salt, 1892). The real breakthrough of an academic animal ethics came in 1971, when Richard Ryder introduced the term "speciesism": a discrimination on the basis of someone's species, by analogy with racism and sexism (Godlovitch & Harris, 1971; Ryder, 1975). The 1970s and 1980s were characterized by the application of different rational¹ theories in normative ethics (mainly utilitarianism and deontological ethics) to animals (Singer, 1975; Clark, 1977; Regan, 1983). In the 1990s, criticism arose from the postmodernist and feminist point of view, against the "cool", rational approach. (Plumwood, 1993; Adams, 1995, 1995b). A new plea for vegetarianism relied on an ethics of care (Adams & Donovan, 1996) or a virtue ethics (Hursthouse, 2000).

Around the turn of the century the debate took a new twist towards (social) psychology and experimental philosophy. Many animal rights ethicists consider the argumentation for animal rights and veganism as solid and completed, but they note that there is more psychology than ethics behind our use of animals (Serpell, 1996; Allen et al. 2000; Joy 2002, 2009; Herzog 2010). The question should be asked why so few people are convinced by logical consistency and rational arguments.

In this dissertation, I will return to the older, rational tradition in animal ethics, the approach of the seventies and eighties. After 40 years of refining the theories of animal ethics, I want to try to present an ambitious, most consistent and coherent ethical

¹ I define 'rationality' in its broadest sense as 'effectivity in means, consistency in ends'. In this context, a rational ethical theory is (very roughly) characterized by an appeal to critical thinking, logic, consistency and reason, using a language of principles, rules or rights. It is distinguished from a more emotional approach to ethics. However, some experimental philosophers (e.g. Greene, 2008) claim that some rationalist ethical principles (e.g. deontological rights) might be the result of underlying intuitive emotional reactions.

system of animal equality. Consistency is the objective. My motivation for returning to this rationalist tradition with its focus on consistency is sixfold.

First, this consistent ethical system of animal equality demonstrates that vegan animal rights people are safe in relying on and following its moral code of veganism. These people can trust the ethical system and don't need to worry that their system contains moral inconsistencies as severe as the inconsistencies encountered in the other, speciesist ethics.

Second, I want to express that critical thinking, consistency and rationality are and should be important elements in our moral lives. Consistency and coherence are strong constraints on ethical systems. These constraints help us to distrust unreliable moral intuitions (moral illusions) and to avoid a hyperrelativistic "anything goes" attitude. Throwing away all inconsistent ethical systems already limits the options to the surviving consistent ethical systems. This makes it easier and more reliable to select the ethical system that best fits our shared and strongest moral intuitions. If ethics would be merely a matter of taste, it should be a matter of consistent taste.

Third, related to the previous point: I believe that a consistent system that best fits our shared strongest moral intuitions has a higher likelihood of compliance. Consistency limits arbitrariness, and a less arbitrary conception of justice might be more politically stable: I believe (and hope) that individuals who grow up in institutions governed by less arbitrary, more consistent conceptions of justice tend to be more motivated to respect its rules (this is an empirical claim, however, that requires empirical evidence).

Fourth, we can try to convince someone to become vegan by appeal to emotion, intuition, empathy or serving a delicious healthy vegan meal. These are all valuable "psychological marketing tricks" in the animal rights movement, and when it comes to encouraging common people to become vegan, these strategies are likely to be more efficient than an appeal to rational consistency. But it is my conviction that becoming vegan for an additional right reason, as a moral duty based on rational, consistent arguments, has some extra intrinsic value.

Fifth, I want to demonstrate that some new things can and should be said in a rational animal rights ethic. In the past 40 years, philosophers underestimated the importance of some problems (such as the predation problem and the sentience problem). I will tackle these problems. And I want to present a clear overview of all basic principles needed in an animal rights ethic, using the metaphor of five fingers of a moral hand.

Sixth, my hope is that people, especially academics and other people interested in ethics and philosophy, would critically consider the arguments in this dissertation and apply them to issues like veganism and the production and consumption of animal products. I will try to demonstrate that the speciesist ethics that a lot of people have, is not internally consistent, and that a feasible consistent theory that respects deep

(phenomenally strong) moral intuitions is possible. Making the ethics consistent results in an ethical system, one of its implications is veganism. For those people who believe in rationality and consistency, these arguments might be an additional motivating factor to become vegan. After critical reflection I believe that veganism is not only a fair, healthy, ecologically sustainable and animal friendly way of living, but it is also a way of living that is most consistent with the deepest moral values of a lot of people.

In order to argue for animal equality, this dissertation contains three parts. The first part refers to the “ethical consistency” in the title, here understood in the broad sense. So I start with meta-ethics, a reflection on the meaning and validity of ethics. In the first part, we will not yet find principles of how we should behave. We first have to set up the rules of the game. If I want to convince a meat eater with rational arguments, we first have to agree on the validity of the rules of the game.

The starting point is our moral intuitions. These intuitions are spontaneous, unreflected gut feelings that something is right, wrong, good or bad. The resulting persistent judgments cannot be justified with further rational arguments. Intuitions are “things that strike us as true without us knowing entirely why they do” (Cohnitz and Häggqvist, 2009, p.3). They arise on their own and are not the result of inferential reasoning.

Moral intuitions have different strengths, where the strength is determined by our willingness (not) to give up the intuition. Hence, the strongest intuitions are the ones that are accompanied with the strongest emotions and the strongest desire to respect them. Therefore, strong moral intuitions are intrinsically motivating, which gives us a first reason to start with those moral intuitions. A second important reason is that meat eaters start with intuitions as well, when they want to justify their meat consumption. So meat eaters and animal ethicists can agree on this part of the rules of the game. I believe that all ethical systems are in the end based on moral intuitions. I hope (and weakly believe) that the strongest moral intuitions that I use in this dissertation to argue for veganism, are shared by meat eaters. In other words: if animal rights activists and meat eaters put all their moral intuitions on the table, selecting the strongest of them to construct a consistent ethical system, we might see that a lot of those intuitions are shared by all parties. This however is an empirical claim that requires further evidence.

After moral intuitions, a second important element is the method that will be used to move towards consistency: universalization. This is the area of reflective ethics. Reflecting on particular situations, some moral intuitions are ignited. We have to translate those moral intuitions into particular ethical rules, applicable to those particular situations. Next, these particular rules should be universalized to all morally similar situations. We arrive at universalized ethical rules.

The importance of this method of universalization is also shared by meat eaters, as we can see in discussions about animal rights and meat consumption. Not only animal rights ethicists, but also a lot of meat eaters give arguments based on universalization. So the good news in convincing meat eaters is that we share the same rules of the game: moral intuitions and universalization. We both value consistency.

Apart from a (hopefully) rather strong consensus on the importance of universalization, this method has another important advantage: it puts strong constraints on the consistency of ethical systems. So the next step is to check whether our universalized ethical rules are consistent in the sense that they form an internally consistent system lacking contradictions. When it comes to the theory of animal equality, we will see that it is not only consistent, but coherent: different intuitions, principles and arguments mutually support each other. They form a web of principles, like a crossword puzzle. As different letters combine into words, different intuitions can be unified in principles. The meat eater will have difficulties trying to show inconsistencies in this coherent structure.

The crossword puzzle analogy helps to clarify the notion of coherence. Coherence, or consistency in the broad sense, has two aspects: non-arbitrariness and internal consistency. Non-arbitrariness in the crossword puzzle means that the letters in neighbouring white boxes should not be random, but should form existing words. Internal consistency means that each white box should not contain more than one letter.

The two rules of the game (one letter per white box and white boxes generate words) place strong restrictions on the possible solutions of a crossword puzzle. Without those two rules, a crossword puzzle will have many equivalent solutions. With the rules, there will only be a few solutions (most of the puzzles have only one, but some might have two or even three parallel solutions). Similarly, non-arbitrariness and internal consistency place strong restrictions on possible ethical systems. As arbitrary and inconsistent systems are thrown away, we are left with only a few possible systems that best match our moral intuitions. As a consequence of these two restrictions, there will be more mutual agreement between the ethical systems that different people have. A consensus will be easier to achieve and there will be less space left for moral relativism.

To illustrate the two rules of internal consistency and non-arbitrariness, we can look at two interesting analogies between atheism and egalitarianism. First, the atheist does not believe in a god that is 1) almighty, 2) all good and 3) allows the evil in the world, because a representation of a god that has those three properties is internally inconsistent. Similarly, an egalitarian animal rights activist does not believe in an ethical system where 1) humans have basic rights, 2) animals can be killed and eaten by humans and 3) speciesist discrimination is not allowed, because an ethical system that contains those three elements is internally inconsistent. As a second analogy, we can say that a theist is in fact an inconsistent atheist: a person who believes in God does not

belief in Zeus, Apollo, Thor, Krishna or any other possible god. S/he is a 99,999...% atheist. This kind of inconsistency refers to the arbitrariness: it is arbitrary to pick out and believe in one of those many possible gods if the amount of evidence for the existence of all those gods is equal (in fact nihil). Similarly, a speciesist meat eater is an inconsistent egalitarian: it is arbitrary to discriminate on the basis of species instead of e.g. race, sex, population, genus, family, order, class or any other possible (biological) category. Picking out the category of species is arbitrary, because the moral relevance of all of those many possible categories is equal (in fact nihil).

We can apply a kind of golden rule of reciprocity to counter arbitrariness and inconsistency. If you may believe in God without evidence, then I may belief in Brahma without evidence. If you may say that we should have blind faith in Allah, then I may say that we should have blind faith in Jupiter. If you may be a speciesist without justification, then I may be a racist without justification. This kind of golden rule of reciprocity is nothing but an application of the method of universalization. As we will see, this universalization puts strong constraints on our beliefs and our ethical systems.

Yet, like theists, speciesists will not easily be convinced by the arguments. The reason is that there are cognitive and moral illusions, in analogy with optical illusions (Purves & Lotto, 2002). We cannot trust all our intuitions. Universalization is a method to find out whether an intuition is an illusion. I will argue as clear as possible how to determine moral illusions. My final goal, in part three of the dissertation, is to show that speciesism based on a moral illusion: it is a stubborn intuition that makes our ethical theory inconsistent.

The principle (or method) of universalism generates a formal principle of equality in terms of impartiality and antidiscrimination. It says that we should treat equals equally in all equal situations. This is a formal principle, because it does not state how we should treat an individual. To give this principle some material content, we have to move to normative ethics. Instead of meta-ethics, normative ethics deal with questions of what is right and good, what should be done and what is valuable.

The second part of this dissertation uses normative ethics to present three material principles of equality. Hence, this part refers to the final word in the subtitle of this dissertation. I will derive three material principles of equality, based on different normative systems (consequentialist ethics, ethics of care and deontological ethics). Together with the formal principle of equality (the impartiality and antidiscrimination principle), these three different material principles of equality can be used to construct a nuanced theory of animal equality. As we will see, these principles of equality are not incompatible with some of our moral intuitions that can be translated into two notions of inequality.

The first material principle of “prioritarian equality” is a consequentialist principle of justice, focusing on just distributions of lifetime well-being. It states that we have to give a strong priority to increasing the lifetime well-being of the worst-off sentient

beings. It can be justified with two different, mutually supporting arguments. One argument is based on a rational thought experiment of impartiality (the veil of ignorance, Rawls, 1971), extended with a high but not maximum risk aversion (need for safety). A second argument is based on a feeling of empathy, extended with a low but not zero need for efficiency.

The second principle of “tolerated choice equality” says that we are allowed to be partial towards those sentient beings with whom we have a personal or special relationship² or for whom we feel a lot of empathy. This partiality is only allowed as long as we could respect similar levels of partiality that all other sentient beings might have. The concern for personal relationships or special feelings of empathy is characteristic of an ethic of care that distances itself from an ethic that is too impartial. I will combine the tolerated choice equality with a well-known principle of equality of opportunity.

The third principle of “basic right equality” brings us to deontological ethics based on duties and rights. The basic right is related to a mere means principle, as it is the right not to be used as merely a means to someone else’s ends. Equality means that all sentient beings have an equal claim to this right. Instead of empathy, this right stems from the feeling of respect. The basic right is coherent with a lot of shared moral intuitions encountered in a lot of different moral dilemmas. An extended version of the mere means principle allows for other deontological intuitions (e.g. the difference between doing and allowing). And more: the extended mere means principle opens space for the tolerated partiality principle that generated the tolerated choice equality. It says that we should tolerate some levels of impartiality. This property nicely demonstrates the coherent interweaving of the different principles.

The three material principles of equality allow us to get a very precise and nuanced picture of discrimination.

The third part of the dissertation is devoted to applied ethics, referring to the “animals” in the subtitle of this dissertation. So we will discuss some issues in animal ethics. First, I demonstrate that speciesism is a moral illusion which results in immoral discrimination. In order to do this, I will present five arguments why the species boundary (the criterion “human”) is not morally relevant, and five other arguments why sentience is morally relevant. These ten arguments cohere with each other and are based on strong moral intuitions and scientific knowledge. One intuition that speciesists have – the prejudicial difference in moral status between humans and non-humans – is not strong enough to overthrow the ten arguments. And if we add some psychological

² This principle is not restricted to a mutually conscious relationship: you might have a personal relationship with an individual that is not consciously aware of being in a special relationship. E.g. the relation with a baby.

insights in the mechanism of discrimination, we get an even stronger case against speciesism.

If we mention sentience³, our next problem is how to know whether a living being is sentient. This sentience problem consists of two parts: first there is the scientific question of the required criteria to test whether someone is sentient. I will briefly present the current scientific consensus on this issue, which roughly says that at least all vertebrate animals with a functioning central nervous system are sentient. The second problem is an ethical one: we now have to do tests to see whether a living being (a fish, an insect,...) is sentient, and those tests might cause pain, fear or distress when the individual is indeed sentient. Are we justified in performing such tests? Are we not using those animals as merely means? I will argue that such tests are not immoral.

In the animal rights discussions over the past 40 years, a highly underestimated problem is the predation problem. As with the sentience problem, I will discuss the predation problem in two parts. First, there is the prey problem: suppose a lion is attacking a zebra. Most people, including animal rights activists, say we do not have a duty to protect the zebra if we could. But if the lion is attacking a human, things change. Isn't this speciesist, and how to reconcile this intuition with an antispeciesist ethics? The principle of tolerated choice equality will solve the issue.

The second issue of the predation problem is the difference problem. What is the difference between a lioness killing a zebra in order to feed it to her two whelps, and a surgeon killing an innocent person in order to use his organs to save two patients in the hospital? In both cases, a sentient being is killed against his will and parts of his body (muscle tissue or organ tissue) are given to other sentient beings in order for them to live. Yet, there appears to be a consensus amongst most animal rights activists, that predation is allowed but organ transplantation is not. So what is the difference between predation and transplantation? How can we reconcile our theory of animal equality with this intuition that there is a difference? In order to solve the issue, I will introduce a new principle, which is in fact based on elements of a carnist ideology. Carnism (Joy, 2002; 2009) is a sub-ideology of speciesism, the opposite of the ideology of veganism. Modern day meat eaters often unconsciously adopt this ideology, justifying their meat

³ Note: I use the words "sentient being", "affective being" and "person" interchangeably. Although feelings and sensations can be neutral, a sentient being has affective (positive and negative) reactions such as liking, disliking, pleasure and displeasure. A person is broadly defined as a being who has personal experiences and preferences. This requires the presence of a perceptual consciousness and the capacity to have positive and negative feelings. When the context makes it clear, "persons" will sometimes refer to "moral agents", a subset of the sentient beings.

consumption by claiming that this behaviour is natural, normal and necessary⁴. By clarifying those three notions, I will argue that violations of the basic right are only allowed when all three criteria are satisfied. So predation is allowed because it is normal, natural and necessary, whereas transplantations are neither normal nor natural. Borrowing aspects from the carnist ideology gives us the advantage that our theory becomes coherent with the intuitions that a lot of meat eaters share.

To make the case for this 3-N-principle even stronger, I relate the 3-N-principle to the intrinsic value of biodiversity. To justify this value, I explore an interesting analogy between two properties: well-being of sentient beings and biodiversity of ecosystems. The intrinsic value of biodiversity introduces an element of environmental ethics, an important branch of applied ethics.

The 3-N-principle (based on the value of biodiversity) can be combined with another principle that can solve the difference problem in the predation problem: the principle of behavioural fairness. Together, we arrive at a fifth principle of equality (a fourth material principle): everyone has an equal right to a behaviour that is both natural, normal and necessary (i.e. a behaviour that strongly contributes to biodiversity). Briefly put: if a zebra is allowed to eat for survival, then so is a lion.

The epilogue of this dissertation presents the metaphor of the moral hand: five basic ethical principles correspond with the five fingers of the moral hand. These five principles are universalism (the thumb), prioritarian justice and the value of well-being in consequentialist ethics (forefinger), the mere means principle and the basic right in deontological ethics (middle finger), naturalness and the value of biodiversity in environmental ethics (ring finger), and tolerated choice partiality and the value of personal relationships in ethics of care (the little finger). These five fingers generate five principles of equality (one formal, four material), respectively: impartiality, prioritarian equality of well-being, basic right equality, naturalistic behavioural fairness and tolerated choice equality. Although these five principles of the moral hand might conflict with each other in particular situations, they can be considered as moral forces that need to be balanced against each other. Just as a physical system with multiple forces (gravity, electromagnetism,...) is not inconsistent, the moral hand is not necessarily an inconsistent ethical system. However, some elements of a speciesist or carnist ethical system are internally inconsistent in the sense that the theory says that something is both allowed and impermissible, without the possibility of balancing different principles.

⁴ Consumption of animal products is not necessary, however, as dietitians claim that a well-planned vegan diet is healthy for everyone, including pregnant women and athletes (ADA, 2009).

In the appendix, I present a review and systematization of the trolley problem, a very famous thought experiment in moral philosophy. A runaway trolley is about to kill five innocent people. Are we allowed to sacrifice another innocent person in order to save the five people?

After the bibliography I present a schematic argument for veganism, clearly indicating all the basic assumptions and logical steps.

Before we dive into the details, I want to give a summary of this dissertation: a structured line of reasoning to arrive at a complete ethical system of animal equality, consisting of clear and coherent universalized ethical principles that best fit our strongest moral intuitions, without too many arbitrary elements.

Summary: towards a coherent theory of animal equality⁵

In this introduction my goal is to construct a coherent ethical system that is capable of dealing with all relevant issues in principle-based animal ethics. The basic line of reasoning of this construction goes as follows: I start with a factual property of the world, which ignites a moral intuition or emotion, i.e. a quick, spontaneous moral response or judgment that has no further rational justification. Then, in a process of reflection, this intuition is translated into a universalized ethical rule, where “universalized” means: “relevant to all morally similar situations”. Sometimes different moral intuitions will mutually support each other, resulting in a set of coherent universalized ethical principles. But sometimes we encounter a new fact or situation that again ignites another moral intuition or emotion, which might be in contradiction with our constructed set of universalized ethical principles. To solve this conflict or moral dilemma, we can either change the ethical principles, or introduce a new ethical principle that trumps the previous ethical principles in that particular situation. This new ethical principle needs to be universalized as well to all relevantly similar situations.

This process continues: we again test the constructed coherent set of universalized principles in new situations, and if we encounter a moral dilemma, we look for further refinements. Eventually, all situations and all facts that ignite moral intuitions should be

⁵ This section is based on Bruers (2014), *Towards a Coherent Theory of Animal Equality*, accepted for publication in *Between the Species*.

covered, and we move to a consistent ethical system of hierarchical universalized principles, where some principles trump others. In other words, we reach a theory in 'reflective equilibrium' (Rawls, 1971), which means that our strongest moral intuitions and ethical principles are coherent (mutually supporting each other).

This approach can be compared with solving a crossword puzzle. The descriptions of the words are the analogues of relevant input data (objective facts in the world as well as moral intuitions that we have). The white boxes refer to the different possible situations and viewpoints, the individual letters represent the intuitive moral judgments from particular viewpoints in particular situations. The words correspond with the universalized ethical principles (applied to all similar situations), and these words mutually support each other and form a coherent solution to the puzzle.⁶ So let's derive a coherent ethic of animal equality, starting from the most basic, indisputable objective facts and moral intuitions.

The construction of a coherent system

Fact 1: All sentient beings have a well-being and they value their own well-being (and everything that contributes to well-being). Sentient beings are beings that have and can subjectively feel interests. They have the experience of having preferences (wanting something). Things subjectively matter to them, meaning first of all that the individual has a mechanism (i.e. a complex functioning nervous system) that enables the individual to have representations of their bodies and environments. These representations can have intentionality, resulting in qualitative experiences (phenomenological sensations or qualia). For example: through my fingers I can feel these pages. I know the difference between this feeling and an absence of feeling, for example when my fingers are anaesthetized. However, just before I paid attention to this feeling of touch, I was not aware of it. There was an unconscious neural activity (but no anaesthesia). Only after I focused on my fingertips, it became a conscious experience or 'quale' of touch. Now, qualia are often neutral. I don't feel an urge to avoid touching paper. But other qualia have valence. They are affective in nature and are evaluated as being positive or negative. A needle in my finger generates a quale that I wish to avoid. This quale is called pain and it generates an urge in me. Once a quale becomes an affective mental state (i.e. a positive or negative feeling or emotion such as pain,

⁶ To be clear, *constructing* an ethical system (based on input data such as moral intuitions generated through thought experiments) is analogous to *solving* a crossword puzzle (based on input data such as descriptions of the words and knowledge of the pattern of the white boxes and the lengths of the words). Hence, constructing an ethical system should not be confused with *constructing* a crossword puzzle.

distress, joy,...), well-being comes into play. These feelings are related to interests, desires or needs: they are nothing but subjective experiences of (un)satisfied interests. Fear, pain and frustration indicate that the needs for respectively safety, bodily integrity and freedom are not satisfied.

Moral intuition 1: Impartiality is morally important. We can consider a two-step process to increase impartiality, from rational egoism to extended contractualism. A rational egoist would strive for a contractarian ethic (cfr. Thomas Hobbes), where all rational beings (i.e. beings with whom one can negotiate) of equal power will become part of the moral community, because those rational egoists gain mutual advantages through the social contract. However, in a first step to extend impartiality, Rawls (1971) used the method of the veil of ignorance to delete the second condition of equality of power. He arrives at a contractualist ethic that also includes rational people in dependent or weaker positions (minorities, future generations,...). The veil of ignorance is a thought experiment, whereby you imagine that you will be born as a rational agent, but you don't know who you will be. You can determine the moral and political laws, based on your knowledge of the natural laws. I would suggest a second step to extend impartiality, whereby we delete the condition of rationality. Imagine that you might be any object or entity in the world, but you don't know who or what you might be. For complete impartiality, you have to imagine you could be a planet, an electron, a pig in the year 3000 or anything you can think of. How would you like that entity to be treated? If you were non-sentient, this question would not matter to you, because nothing done to you will influence your well-being. You would not have a well-being, experiences or preferences. The kind of treatment becomes important only for those beings whose well-being can be influenced by moral agents. Non-sentient entities should not be taken into account in this moral evaluation. So the least arbitrary and most impartial thing to do is to delete both conditions (of rationality and equality of power), which is what Rowlands (1998) argued, from which it follows that well-being still remains important.

Universal ethical principle 1: All moral agents should strive towards impartiality in all situations, and should take everyone's well-being into consideration in an impartial way. Moral agents are people who are able to understand the notion of impartiality.

Fact 2: Empathy is meaningful for all and only for sentient beings (feeling empathy for non-sentient beings such as teddy bears would be a kind of projection of emotions). Empathy is the capacity to experience or sample the emotions of others. This emotional response occurs when the perspective (frame of reference) of the other is taken.

Moral intuition 2: Compassion (empathy plus the desire to alleviate the suffering of the other) is a moral virtue.

Universal ethical principle 2: All moral agents should develop compassion in all situations (hence also towards all sentient beings). Moral agents are people who are able

to develop compassion, are able to understand the virtue of compassion, and are able to help others. Those moral agents should try to improve the well-being of others.

The above two universal ethical principles are coherent with each other, and give a rational and emotional basis of the moral importance of sentience. They are based on contractualism, consequentialism and virtue ethics. The coherence gets even stronger when we consider the following two moral intuitions. A) Mental capacities (self-consciousness, rationality,...) are morally important. They are very special, complex and vulnerable, hence worth protecting. B) Babies and mentally disabled humans have rights because they have something morally important. They have a higher moral status than human egg cells, skin cells, dead human bodies, plants or stones. Together with the fact that sentience is the only mental capacity that mentally disabled persons have in common with other humans who have strong rights, A and B generate two extra reasons why sentience is important. Furthermore, the link between rights and sentience is also not farfetched: rights protect interests; feelings detect interests.

This gives us a strong coherent case for the moral relevance of sentience. It is a scientific question (i.e. a matter of fact) what entity has a well-being and how its well-being can be influenced. We can briefly compare this moral relevance of sentience with the moral irrelevance of a criterion such as the species *Homo sapiens*. First, the species is one of the many biological classifications, thus it is arbitrary to pick a specific species and not a specific population, genus, family, order, class,... Second, the definition of a species is very complicated. One of the definitions refers to a set of individuals who could get fertile offspring. But reference to fertility and offspring is very artificial and farfetched when it comes to determining who has rights. Third, science will never be able to determine whether a human-chimpanzee hybrid, a human-animal chimera, an ancestor (*Australopithecus*, *Homo habilis*,...) or a genetically modified humanoid should still be called *Homo sapiens*. The boundaries are fuzzy. Fourth, all species are temporally related to all other species in a similar way, as populations can be spatially related in a ring species (a ring species consists of a spatial spreading of populations, where A can get fertile offspring with B, B with C, but C not with A). Fifth, if the moral status of a species is determined by genes or bodily appearance, then it is also very arbitrary to pick out those genes or bodily characteristics and not others (such as skin colour). We are not responsible for our genes, so it would be a violation of the desert principle if we based moral status on genes. In summary, the species boundary is too arbitrary, artificial and abstract to be morally relevant.

So far, our ethic is not yet unambiguous and clear. We observe that there are different sentient beings and multiple ways to influence their well-being (for example: increasing everyone's well-being a little bit versus increasing the well-being of one individual a lot). So what is a just distribution of well-being? First of all, we value parsimony and simplicity. One simple solution would be to add the levels of well-being of all sentient beings for a specific time interval, and then take the sum over all times.

Then we could try to maximize this sum. This is sum-utilitarianism. But there are also other simple options, such as trying to maximize the well-being of the worst-off sentient being (the one with the lowest level of well-being). This is maximin-utilitarianism. However, according to many people, both sum-utilitarianism and maximin-utilitarianism have some counterintuitive implications. With sum-utilitarianism, it is morally good to sacrifice one individual in order to increase the well-being of others, or to kill one individual and replace him with another sentient being, or to keep on breeding sentient beings in order to increase the sum of well-being. The latter is known as the ‘repugnant conclusion’ (Parfit, 1984): an overpopulated world with a trillion individuals with a well-being slightly above zero, might be better than a world with only a thousand individuals who have a satisfyingly high level of well-being. Our moral intuitions go against these conclusions. These conclusions can be avoided by introducing a level of risk aversion.

Fact 3: There are many sentient beings, and some beings can be worse-off than others. This fact implies that from behind the impartial veil of ignorance, how to maximize your well-being becomes a game of chance. Mathematically, sum-utilitarianism implies that the expectation value of your well-being will be maximized. But you have to be aware that there is a risk that you might be born as one of the worst-off individuals. For example: two individuals might have well-being levels equal to 10 and 100, so the expectation value will be equal to 55 (the average). In sum-utilitarianism, this situation would be equal to the situation where those two beings both have a well-being of 55. The problem is that in the first situation, you might end up as the person with level 10. When much is at stake, most moral agents have a risk aversion (need for safety – to play it safe), and in this game of chance, this means that they would not opt for sum-utilitarianism, but to some kind of prioritarianism: giving priority to increases of well-being of the worst-off positions. Therefore they prefer the second situation (with equality of well-being). If you have maximum risk aversion (a maximum need for safety), you would take the maximin-utilitarian strategy (maximizing the minimum/lowest well-being), giving all priority to the worst-off position, because you are so worried at becoming this worst-off individual. If you have zero risk aversion, you are a sum-utilitarian. A high but not maximum level of risk aversion would result in a prioritarianism that is in between maximin-utilitarianism and sum-utilitarianism. We could call this ‘quasi-maximin prioritarianism’.

Moral intuition 3: A (high) level of risk aversion is good (especially when much of your well-being is at stake; then most people are risk averse).

Universal ethical principle 3: Quasi-maximin prioritarianism should be applied in all situations. Mathematically, this principle can be expressed as the maximization of a power average of values of life of all sentient beings. The power in the power averaging corresponds with the level of risk aversion behind the veil of ignorance. The value of life (lifetime well-being) refers to the total preferred well-being of an individual over

his/her complete lifespan (that spans from the first till the last subjective feeling of the individual). This preferred well-being is the value that one would ascribe to living the complete life of that individual, when looking from the most impartial point of view, e.g. from behind a veil of ignorance. The value of life contains everything that would matter to you, everything that would be valuable to you, all the preferences that you would have, if you would live the life of that sentient being.

Quasi-maximin prioritarianism has some elegant features. It avoids the abovementioned objections against sum-utilitarianism, and also a lot of objections against animal ethics. First, consider the idea of painlessly killing someone (for example in his sleep). From behind the veil of ignorance, you cannot prefer such killing, even if you are not aware that you will be killed. This means that a sentient being should now be defined as a being that has already developed the capacity to feel and has not yet permanently lost this capacity. Indeed, value of life starts from the first feeling and ends at the last feeling.

Next, take the problem of replaceability. Is it allowed to kill a sentient being (painlessly), and then let another sentient being be born? This happens when we breed and slaughter cows. If we kill a sentient being, his value of life will be e.g. 5, whereas it would have been 10 otherwise (when he lives a full life). So in a first option, one individual will have a life with total well-being equal to 5 (an early death), and a second one will also have a short life with total well-being 5. In a second option, we will have only one being, with level 10 (a full life). From behind the veil of ignorance, in the first option you will get a low value of life equal to 5. In the second option, you are sure you will have level 10. A sum-utilitarian would say that the both options are equal, because the total value of life equals 10 in both situations. But I would prefer the second situation, and that's also what our prioritarian theory says, because this theory uses a (power) average. Therefore, sentient beings are not replaceable. Also the repugnant conclusion (the idea to keep on breeding sentient beings until their values of life are about to drop below zero), can be avoided, by simply noting that behind the veil of ignorance you would not prefer an overpopulated world where everyone has a very low value of life. So quasi-maximin prioritarianism avoids the often heard argument that breeding livestock animals is good, because they owe their lives to the breeders, and it is better to live a life on a farm than not to be born at all (this might not be the case for animals living on a 'factory farm'). According to our prioritarianism, the choice is not between an existing life on a farm versus a non-existing life, because as said above: in each choice, we only consider the sentient beings that exist in that world-history.

Another famous problem in animal ethics is the lifeboat dilemma (e.g. Regan, 1983). Suppose there are different sentient beings in a lifeboat, but we cannot save everyone. Those beings can have different expected life expectancies, but they can also differ in complexity (richness) of emotions, the amounts of needs, the levels of satisfaction when needs are satisfied,.... This means that the potential values of life can differ amongst the

different sentient beings in the lifeboat. The potential values of life between a (mentally normal) human, a dog or a frog can differ. This influences our choices whom to rescue. As Regan argued, it might be required to sacrifice the dogs first, because they experience a less rich life than the humans. However, Regan said that the life of one human would trump the lives of a million and more dogs. According to our prioritarianism (the veil of ignorance with a high but not maximum level of risk aversion), there would be a number of dogs, above which the loss of that amount of dogs would be worse than the loss of one human life.

The quasi-maximin principle is coherent with a lot of our moral intuitions. And there is a second way to arrive at this principle.

Fact 4: There might be situations where we can decrease someone's well-being with a huge amount (e.g. drive him/her into extreme poverty) in order to increase the worst-off position with a negligible small amount.

Moral intuition 4: Efficiency is important to some degree. Empathy might have a tendency to give absolute priority to improving the worst-off individual, which results in a maximin strategy. But if we value efficiency, we would not sacrifice too much well-being.

Universal ethical principle 4: This equals quasi-maximin prioritarianism (principle 3). We should maximize the value of life of all sentient beings, giving a strong priority to increase the lowest values of well-being. In other words: we should maximize the value of life of the worst-off individuals, unless this is at the expense of much more well-being of others.

In summary: a rational approach of impartiality (the veil of ignorance) with a high but not maximum risk aversion (need for safety) coheres with an emotional approach of compassion with a low but non-zero need for efficiency. The two approaches represent two different points of view: the rational approach looks at a situation from the outside, from an impartial point of view behind a veil of ignorance. The emotional approach is more down to earth: it looks at a situation from the inside, from the subjective experience of compassion with others. These are two approaches resulting in the same quasi-maximin prioritarian principle.

This principle has two disadvantages. As a first problem, the values of life are very difficult to measure and compare. All we have is our empathy, our scientific knowledge and our imagination. We have to try placing ourselves in the position of others, by using empathy, or by imagining that we could be the other individual, with all his or her needs and feelings. So the 'emotional' method of empathy and the 'rational' method of the veil of ignorance are actually two rules of thumb to make educated guesses about the order of the values of life of different individuals. Empathy and imagination are virtues to be developed and already allow us to move quite far.

A second disadvantage is that the level of priority given to the worst-off (in other words: the level of risk aversion or the need for efficiency), is in some sense arbitrary.

The level is somewhere between 0 (sum-utilitarianism with zero risk aversion) and infinity (maximin-utilitarianism with maximum risk aversion). However, I believe our coherent picture is strong enough to withstand this objection. The arbitrariness is less bad than overriding a coherent set of strong moral intuitions. The good thing is that no-one has a strong preference to a sharp level of priority. No-one says the value should be 748. It's more like a fuzzy range that we prefer. So we can and should be a bit tolerant to the levels of priority that other moral agents would prefer, and this means we can be flexible and could come to a democratic or mutual consensus between all moral agents. But once we have set a level of priority, we should apply it consistently in all relevantly similar cases.

The quasi-maximin prioritarianism is the basic framework of a coherent ethical system of animal equality. All sentient beings are in some sense equal from an impartial perspective such as behind a veil of ignorance. It is a consequentialist ethic, because it only looks at outcomes of values of life. Giving a level of priority for the worst-off positions, some people (true consequentialists) might prefer to stop the construction of a coherent ethical system here. However, there are some more intuitions that do not fit in the prioritarian ethic. We first discuss an intuition related to an ethic of care and next an intuition related to an ethic of rights.

Fact 5: There is a possible situation where I have to choose between a sentient being I hold dear and one or more other unknown sentient beings. E.g. in a burning house dilemma, where I have to choose between saving my child or other individuals from the flames.

Moral intuition 5: I am allowed to help the person I hold dear.

Universal ethical principle 5: It is allowed to be partial in all situations of aid where someone is involved whom you hold dear (with whom you have a personal relationship or strong feelings of empathy), as long as we tolerate similar levels of partiality of everyone else. This principle of tolerated partiality trumps the above prioritarian principle to some degree, but not too much.

Burning house dilemmas such as “Your child or the dog?” (Francione, 2000) are often used to criticize animal equality. But here I introduce a new principle of tolerated partiality, which hides a new kind of equality: tolerated choice equality. In the burning house, I would save my child instead of someone else, which points at an emotional inequality in favour of my child. But I can still consider all individuals in the house as being equal in some other subtle sense, if for example I tolerate your choice to save someone else instead of my child. A white racist would say that it is immoral to save black children from the house instead of white children. A speciesist would say that it is immoral to save someone belonging to another species. But if someone has an

emotional connection with a dog, the principle of tolerated choice equality says we should tolerate his choice to save the dog. Saving a dog instead of a human⁷, saving a mentally disabled orphan instead of a mentally normal child, or saving your lover instead of two unknown persons, might be violations of the quasi-maximin prioritarian principle. But I think we are allowed to violate this quasi-maximin principle to some degree. Also here we could try to reach a democratic or mutual consensus between all moral agents, about the degree of violation that is allowed. We should apply this degree of partiality consistently in all situations.

Fact 6: The organ transplantation problem. There is a possible situation, where five patients in a hospital would die unless we sacrifice an innocent person against his will and use five of his organs for transplantations. This would be allowed according to prioritarianism, because the (power average) lifetime well-being would be higher if the innocent person is sacrificed.

Moral intuition 6: I (and most people) feel emotional distress and restraint to sacrifice this one person against his will. We should not sacrifice someone, even if prioritarianism is violated and even if someone I hold dear is one of the patients in the hospital. So this intuition trumps both prioritarianism and tolerated partiality.

There are a lot of other moral dilemmas where we can use someone without his/her consent as merely means to save others. Torturing someone in order to gain information about a bomb, throwing someone (a sentient being such as a mentally disabled human) in front of a runaway trolley in order to block the trolley that is about to kill other people, using someone as a shield against bullets, using someone as a slave, using someone in medical experiments, using someone as a scapegoat to stop a riot, terror bombing civilians in order to demoralize the enemy, raping someone, killing and eating someone (cannibalism), trafficking,... All these situations generate moral intuitions that are very coherent if we translate them into the following deontological principle (an interpretation of a Kantian ethics).

Universal ethical principle 6: All sentient beings have a basic right not to be used as merely a means to someone else's ends. A victim is used as merely a means, when two conditions are met. 1) A moral agent causes the victim a 'disrespectful harm' against its will: the victim has to do or undergo something that s/he does not want. Examples of disrespectful harm are a treatment as property or commodity (see Francione, 2000) or a violation of bodily integrity. 2) The presence of the body of the victim is required in order to reach the ends. For example without the body of an animal, we could not produce an animal product (meat, eggs,...) for consumption. The latter is an important criterion because there are moral dilemmas whereby you are allowed to cause harm to

⁷ I tolerate that you give more food and medical assistance to your pet than to a hungry child far away.

someone in order to save others (for example redirecting a threat towards one person in order to save a group of people). In those dilemmas, the presence of the victim was not required in order to save the others.

This principle is coherent with the notion of respect, which is next to empathy an important moral virtue, and it is coherent with the notion of intrinsic value (the opposite of instrumental value) as well.

The ethical principle of the basic right trumps both the principle of priority and the principle of tolerated partiality. But the basic right is not absolute: if the principle of priority is strongly violated (if thousands of sentient beings will die), then a basic right might be violated (this corresponds with a need for efficiency). As with the above principles, this level of violation can be determined on the basis of a democratic or mutual consensus among moral agents. And here we have flexibility as well: there are different levels of harm, there is a morally relevant gradation in someone's ends (from the vital needs of many sentient beings to the luxury ends of one individual), and there is a gradation in the level of sentience and mental capacities. These gradations could be coupled. For example: a being with higher levels of morally relevant mental capacities has a stronger claim to this basic right.

Let's briefly apply this principle to the 'least harm' objection against veganism (Davis, 2003). Suppose that a meat eater can kill and eat one cow, whereas a vegan needs a crop field to get the same amount of nutrients. Suppose using that crop field accidentally kills five mice. The meat eater causes least harm, but s/he violates the basic right of the cow, which is worse. The mice are not used as merely means, so therefore veganism remains the morally better choice. (For further criticism on the least harm argument of Davis, see Matheny, 2003, and Lamey, 2007).

We now arrive at an ethical system with three principles of equality. The first is based on impartiality (interchangeability of sentient beings) and results in a form of prioritarianism. According to this theory, if we have to choose between two situations that have equal total well-being, we should choose the one with the most equal distribution of well-being. The second is a tolerated partiality, whereby we tolerate the choices of others to save those they prefer. From this tolerated partiality, the individuals in a burning house inherit a 'tolerated choice equality'. This principle weakly trumps the first principle. The third principle is a basic right equality, and this trumps the two former principles to a strong but not absolute degree. All beings with similar levels of the relevant mental capacities have an equal claim to the basic right not to be used as merely a means to someone else's ends. The three principles are related to, respectively, a consequentialist ethic of well-being and justice, a feminist ethic of care and a deontological ethic of rights.

These three principles imply veganism. Consider a dairy cow in the livestock industry and a human who likes to eat cheese. Start with the veil of ignorance. In one situation, dairy cows are not bred, so we can only be a human being, who has a value of life equal

to 10. In the second situation, this human enjoys the cheese (his value of life increases to 11), but the cow has a miserable life (suffering in the livestock industry, early death,...). So her value of life equals 3. According to quasi-maximin prioritarianism, the first situation is preferred. If you'd choose the second situation, from behind the veil of ignorance, you have probability $\frac{1}{2}$ to end up in the worst-off position. (According to sum-utilitarianism, the second situation is better). Tolerated partiality is also violated: if we prefer the enjoyment of cheese above the use of the cow, we should also tolerate the other option: breeding women and using their breast milk to make cheese for cows (suppose the cow likes human cheese). This we would not tolerate. The third principle is also violated, because the cow in the livestock industry is used as merely a means (her bodily integrity is violated and she is treated as property).

With these three principles, we arrive at a coherent system that best fits our strongest moral intuitions. Some intuitions based on speciesist judgments are not compatible with this system of animal equality. These intuitions are too weak and cannot be incorporated without introducing highly arbitrary and artificial constructions, so we have to dismiss these speciesist intuitions as being moral illusions. Although our theory implies veganism, it still allows for some partiality (the tolerated partiality meets our intuitive preference for some individuals). However, there is one serious problem remaining.

Fact 7: Obligate predators need meat in order to survive. If obligate predators cannot use other sentient beings as merely means, they will all become extinct. If principles 4, 5 and 6 are universalized to predator animals, this would imply that they have to become extinct.

Moral intuition 7: Obligate predators are allowed to hunt and hence violate the basic rights and well-being of prey. It would be a tragedy if they became extinct.

It is not easy to formulate a clear principle that is coherent with this intuition as well as with the intuitions that we encountered before. If we suppose that biodiversity has a moral value, then we have the following option.

Universal ethical principle 7: If a sufficiently large group of sentient beings became, by an evolutionary process, dependent on the use of other sentient beings for their survival, they are allowed to use other sentient beings for that purpose (until feasible alternatives, that don't violate basic rights, are found⁸).

If we suppose that biodiversity has moral (intrinsic) value, and if we define biodiversity as the diversity of everything that is the direct product of evolutionary

⁸ Such alternatives could be the production of artificial (cultured) meat to feed the predators, genetically or psychologically reprogramming predators to change their behavior, the use of wildlife contraception to control prey populations,...

processes, then this seventh principle becomes coherent with the value of biodiversity. So the existence of predator animals contributes to biodiversity and we should not destroy that biodiversity.

This principle is also coherent with a 'triple-N-principle', which refers to the three values 'natural, normal and necessary' of a carnist ideology (Joy, 2009). This connection works if we define natural as: behaviour that is a direct consequence of a process of evolution (genetic mutation and natural selection). So it refers to an 'evolutionary process'. Normal means that the behaviour happens a lot, so it refers to a 'sufficiently large group'. And necessary means that those beings would die if they no longer exhibit that behaviour. This refers to 'dependency for survival'.

Putting the three criteria together, natural+normal+necessary means that a lot of biodiversity would be lost when the behaviour stopped. And a lot of biodiversity has a lot of moral value; sufficiently enough to trump the basic right and well-being of prey animals. Predation is normal, natural and necessary, so it is allowed (as long as there are no feasible alternatives), even if it violates the basic right. For humans, eating animal products is not necessary (according to the Academy of Nutrition and Dietetics (ADA, 2009)), so we are not allowed to violate the basic rights of animals. Organ transplantation (by sacrificing a sentient being against his will) is not allowed either, because it is a violation of the basic right and it is not normal and natural (although it is necessary for the patients).

Note that this value-of-biodiversity principle is completely unrelated to the value-of-sentience principles discussed before, although we could compare biodiversity as an intrinsically valuable property of ecosystems with well-being as an intrinsically valuable property of sentient beings. Both ecosystems and sentient beings are unique and irreplaceable entities that have a tendency to increase their corresponding valuable properties (biodiversity and well-being). In itself, the biodiversity principle seems arbitrary, but it is coherent with a lot of moral intuitions that a lot of people share. For example: moving around and killing insects (by accident) is considered allowed, even if scientists are able to demonstrate that insects are sentient. But the 3-N-principle says that moving around is natural, normal and necessary behaviour of animals. The same goes for procreation, even if the animal species does not sufficiently contribute to the (power) average well-being of a prioritarian theory. Procreation is natural, normal and necessary, and a lot of biodiversity will get lost if some species were not allowed to procreate.

The 3-N-principle, based on the value of biodiversity, generates a fourth principle of equality: naturalistic behavioural fairness: all natural beings (who contribute equally to biodiversity) have an equal right to a behavior that is both natural, normal and necessary (i.e. a behavior that contributes to biodiversity). Natural beings are those beings that originated by natural evolution.

Finally, we also have situations where predators attack us or beings that we hold dear. Our intuition says we are allowed to defend ourselves and others, and we have a stronger duty to protect some individuals with whom we have special relationships. All sentient beings have the right to defend themselves or others, they have the right to be partial in such decisions, as long as they respect similar levels of partiality of others (see principle 5) and as long as biodiversity is not threatened. If we wish, we could also add that we have a duty to protect all beings who have (or will develop) moral agency or rationality. Those rational beings not only feel their interests, but they also know and understand their interests. This rationality applies to most human beings, except e.g. seriously mentally disabled human orphans. This satisfies people's intuitions that we have a duty to protect humans from predators. (But if we say that we have a duty to protect mentally disabled humans whereas we do not have a duty to protect non-human animals, because all humans have a higher moral status than non-humans, then we become too partial. This kind of speciesism, like racism or sexism, is a kind of partiality and arbitrariness that we cannot tolerate.)

This completes the process. We now have a theory of animal equality, with clear and coherent universalized ethical principles that best fit our strongest moral intuitions, and without too many arbitrary elements. In the epilogue of this dissertation, I will relate the above seven universal ethical principles to five principles of the moral hand. Universal ethical principles 1 to 4 are unified in a forefinger principle of justice and the value of lifetime well-being. Universal ethical principle 5 corresponds with the little finger principle of tolerated partiality. Universal ethical principle 6 corresponds with a middle-finger mere means principle and the basic right to bodily autonomy. Universal ethical principle 7 corresponds with the ring finger principle of naturalness and the value of biodiversity. Finally, the method of translating particular moral intuitions into universalized ethical principles corresponds with a formal thumb principle of rule universalism.

Part 1 Ethical consistency

Chapter 1 The basic elements

In this section I present the basic elements in a construction of a coherent ethical system. Those elements consist of the input data (moral intuitions), the method (the principle of universalization) and the goal: an ethical system which is internally consistent (contains no contradictions), clear (contains exact, unambiguous formulations), parsimonious (does not contain unnecessary or arbitrary elements) and comprehensive (compatible with as much as possible of the strongest moral intuitions, applicable to all situations). I will defend this approach by referring to several analogies borrowed from the empirical sciences, mathematics, geometry, grammar, taste evaluation and crossword puzzles. There is however one serious issue in this approach: the existence of illusions. These are the pitfalls: we cannot trust all our input data (our moral intuitions).

1.1 The input data: moral intuitions

Intuitions can be roughly described as immediate, automatic, fast, non-inferred, a priori, spontaneous judgments (or beliefs) that lack further justification; the typical gut feelings, or responses of a system in ‘automatic mode’. It can be compared with e.g. perception or aesthetic judgments. In morality, I consider the lack of further justification, or the fact that we cannot give rational arguments to justify intuitions, as the relevant property of moral intuitions. Morality is based on moral intuitions about what is right and what is good. These intuitions are judgments with a motivational, prescriptive and often emotional content. We have a desire to comply with our moral intuitions. The stronger a moral intuition, the less willing we are to act against it (or

tolerate someone acting against it).¹ Morality is ‘emotionally driven’, based on emotions such as empathy and indignation. The motivational component means that you yourself feel the urge to do something or refrain from doing something, and the prescriptive component means that you want others to do similar things. Morality is non-cognitive: it does not contain statements with a truth value. It rather expresses emotions, attitudes and prescriptions.

The counterpart of intuitions is reflective thinking. This is the area of ethics, to be distinguished from morality.² Ethics moves beyond the unreflective intuitions of morality: it is characterized by a search for principles as basic building blocks (axioms) of an ethical system. These ethical principles are reflective and clear expressions of underlying moral intuitions.

Although moral intuitions often appear in a blink of an eye, it is not self-evident to articulate all your moral intuitions. In order to interrogate your morality, we can use special devices or thought experiments. These are moral dilemmas developed to test or discover new intuitions. One example of a moral dilemma is a trolley dilemma (see also appendix for a review of the trolley problem): suppose five people are on a track,

¹ Some moral intuitions are more emotional in nature (e.g. the feeling of disgust about unharmed, safe, consenting incest between siblings). Others are more ‘perceptual’ or ‘rational’ in nature (e.g. the spontaneous judgment that saving more lives is better than saving less). Some intuitions are inborn (e.g. “Don’t kill your children”), others are learned or acquired (e.g. some racist intuitions about black people by a white supremacist). Some intuitions arise in concrete situations (e.g. “Don’t push the fat man in front of the trolley”), others might be more abstract or formal (e.g. “Treat equals equally”). In this dissertation, I consider those distinctions as less relevant, because the goal is a ‘reflective equilibrium’ (Rawls, 1971) where our moral intuitions are brought into a coherent system of ethical principles. In the movement towards reflective equilibrium, some intuitions need to be revised or disposed. In that sense, it doesn’t matter whether the intuition is more emotional or rational. What matters is the perceived strength of the intuitions and our willingness to revise or trump the intuitions: which moral intuitions are more revisable or surpassable than others, which intuitions are stronger to resist modification, which intuitions are stronger to surpass others? I do not think that there is a clear correlation between e.g. revisability and the degree of emotionality (versus rationality) of intuitions. Some intuitions that tend to be more emotional in nature might be stronger than other more rational intuitions. And some intuitions on the rational side of the spectrum might easily override some emotional intuitions. Neither do I see a clear correlation between strength of an intuition and the degree of innateness. Therefore, I am not much concerned about the degree of emotionality or innateness of moral intuitions. I am rather concerned about our willingness to revise intuitions and the extent to which intuitions can be brought into a coherent ethical system. I agree with Rawls, who recognizes emotionality (e.g. when we are upset or frightened) as an error-disposed condition of moral intuitions (Rawls, 1971, p41; Brophy, 2009, p13), but some kinds of emotionality (e.g. empathy) might also increase the credibility of intuitions (Brophy, 2009, p115).

² Some animals such as chimpanzees, dolphins and dogs have a (proto)morality (Schermer, 2004, p.16; Bekoff & Pierce, 2009): they can feel empathy, they can cooperate, they demonstrate altruistic behavior or they have intuitions of fairness. It requires more complex rational thinking to move from unreflective (intuitive) morality to reflective ethics.

unaware of the oncoming trolley. They will all die if you do nothing, because you see that the trolley driver is unconscious. You are standing on a bridge, and next to you, exactly above the rails, is a really heavy man. You can save the lives of the five people if you push the man from the bridge, because he is heavy enough to block the trolley. Most people (men and women, from different cultural backgrounds) intuit that it is impermissible to push the heavy man.

In this dissertation we will encounter other moral dilemmas. Such dilemmas often appear to be farfetched or unrealistic, but remember that in order to discover basic laws of physics, scientists agree that performing special experiments (e.g. using particle accelerators) in a thoroughly controlled manner (e.g. in a vacuum) is the best strategy to interrogate nature. Even if those experimental set-ups appear to be everything but similar to the world we experience, they are very instructive to search for physical laws. The idea behind those ‘exotic’ experimental set-ups is to exclude disturbing factors from the experiment. I believe the same goes for ethics: exotic thought experiments like the trolley dilemma can be fine tuned to interrogate our intuitions, eliminating disturbing elements.

Now that we have encountered a moral intuition, it appears that, for a lot of people, it is difficult to translate or express that moral intuition into ethical principles. Why do we let five people die if we could save them? One principle might be: don’t kill. But that’s too vague. A more accurate formulation could be: do not act if action results in the death of a person who would not have died otherwise.

Once we have such hypothetical ethical principles, we have to test whether the resulting ethical system is consistent and whether the principles are both internally consistent, as well as consistent with other moral intuitions. Consider another trolley problem: again five people are on a main track. If you do nothing, they will die. But this time you could turn a switch so that the trolley will take a side track. Unfortunately on the side track one person will be killed. The structure is similar: doing nothing means five people die, acting means five people are saved and another person dies. Most people intuit that we are permitted to act. This seems to be in contradiction with our hypothetical ethical principle, so either we dismiss one of the intuitions, or we refine the ethical principle. In later sections we will discuss in more detail this reflective process of looking for a consistent system.

I believe that moral intuitions are a very important element in ethics and a valid starting point of deriving an ethical system of animal equality. Some ethicists (e.g. some utilitarians such as Singer, 2005) might claim that their ethical systems are completely detached from any moral intuitions. However, all systems start with axioms that can be considered as intuitions. Consider sum-utilitarianism. For those utilitarians it seems self-evident that we should look at consequences and that we should maximize a property. But why are only consequences important, why should we focus at the maximization of something, and what should we maximize? For those utilitarians it

seems self-evident that we should maximize that thing that people want to maximize, such as well-being. But how do we aggregate well-being? For those utilitarians it seems self-evident that we should take an aggregation that reflects impartiality. But what impartial aggregation should we take? For those utilitarians it seems self-evident that we should take a simple aggregation formula. But why should we value simplicity and what simple formula should we take? For those utilitarians it seems self-evident that we should take the sum of well-being. But why should we maximize the sum of utilities instead of e.g. the product, which is equally simple from a mathematical point of view? In summary, even utilitarians are faced with several self-evident beliefs that they are not able to justify any further.

Also virtue ethics and deontologists are based on the intuitions that virtues (beneficence, compassion, honesty,...) or some other properties (intentions,...) are important. All ethicists use intuitions.³ And in the context of animal rights, meat eaters have and live by moral intuitions as well. They have the intuition that personal liberty and choice of food are important, that (some) animals have lower moral status that allows humans to eat them... All moral agents have moral intuitions and they are not easily tempted to do something against their moral intuitions. So, respecting moral intuitions is important. When I derive an ethical system and I want to convince meat eaters to become vegans, I should propose a system that is highly compatible with the strongest moral intuitions that both I and those meat eaters share. The strength of a moral intuition is inversely related to the willingness to override the intuition. Hence, the strongest moral intuitions are the most motivating, because they generate the strongest desires to respect them.

We cannot escape the idea that moral intuitions are the only input that we have in ethics, and that moral agents have difficulties in overriding their strongest moral intuitions. So we should cherish those strong moral intuitions. Ethics without moral intuitions is like science without experimental facts. In all aspects, from science to daily life, we need some relevant input. Let's consider the following six examples that can be used as analogies of ethics.

³ Perhaps one could make distinction between 'cool' (more cognitive) intuitions and 'hot' (more emotional) intuitions. For example Greene (2008) criticized a lot of deontological judgments, claiming that they are based on unreliable, 'hot' or alarmlike emotional responses (such as disgust). If such distinction is possible, we can construct an ethical system based on only the 'cool' intuitions. But for me it seems difficult to clearly make this distinction (perhaps neuroscience can help in making this distinction?). And it is not clear why all 'cool' and none of the 'hot' intuitions are reliable. Therefore, I will make another distinction, between those intuitions that fit in a (strong) coherent framework and those that don't.

1) Physics and other empirical sciences. In the field of scientific research as a cognitive activity, we have observations as input data. These observations are very specific, for example: at time T at place P under conditions C, I saw object O falling. In general, the information about the external world is received through sensory data, perceptions by one of our five senses.

2) Mathematics and algebra. Here we start with a priori knowledge. Three is a number and four is one higher. You have to accept this intuition, otherwise we can't move on in mathematics.

3) Geometry. This line segment is similar to that. There is some property (e.g. length) about lines that make them similar. That is an intuition in geometry.

4) Grammar. According to Chomsky (1986), people from different countries have an inborn faculty of language in their brains, which means that they have intuitive judgments about the grammatical correctness of a sentence. "The gnorfl is sprinkle" is a good grammatical sentence, I just know it, even if the content is meaningless. The grammatical intuitions are interesting to better understand how morality functions. John Rawls (1971) and others (Hauser et al. 2008) proposed an analogy between our language faculty and our moral faculty. The language faculty generates intuitive grammatical judgments about the grammatical correctness of sentences, just like the moral faculty generates intuitive moral judgments about the moral goodness of situations. We intuitively see that an act is morally right and a sentence is grammatically correct. And as with morality, people often have difficulties in expressing why a sentence is grammatically right or wrong; they just know it. It takes some effort to look for the principle that expresses the specific intuition. As with grammar (Chomsky, 1986), some moral intuitions and lines of reasoning are universal (independent from culture) and likely inborn. Thus we can speak of a universal moral grammar (Mikhail, 2000; 2007; O'Neill & Petrinovich, 1998).

5) Crossword puzzles. The input data are the descriptions and number of white boxes for the words.

6) Taste evaluations and aesthetic judgments. The input data are taste preferences of specific products. It appears that people from different cultures like some product made from cane sugar.

In all those areas we have input data, be it evaluations, judgments, observations, given information,... Moral intuitions can be compared with those input data. The above six analogies give us a nuanced picture about the role of intuitions in our morality. I explore further these same six analogies below, in the discussion of the method.

1.2 The method: rule universalism

The method to be used in order to set up an ethical system is universalization. First the moral intuition is expressed as a particular ethical rule, valid for that particular situation. Universalization then consists in extending this particular rule to all other similar situations. In brief: “Equal moral judgments for all morally similar situations.” A situation is morally similar to another if it is similar with respect to all morally relevant characteristics of those situations. This is a formal method: it does not have material content, i.e. it does not state what characteristics are morally relevant.

The idea of universalization is mostly elaborated by ethicists Immanuel Kant (1785) and R.M. Hare (1991). The most general expression of the principle of universalism reads: “You must (may) follow the rule that *everyone who is capable, rational and informed* must (may) follow *in all morally similar ways* in *all morally similar situations* towards *all morally equal individuals*.” The four parts in italics in this expression point at four kinds of universalism: the moral agents (the actors who are capable of doing something), the moral patients (the receivers of a benefit or harm, as morally equal individuals), the acts (the similar ways of doing something) and the situations. The moral agent A does (or refrains from doing) an action C to moral patient P in situation S.

The principles of impartiality and anti-discrimination are clearly universalizations with respect to patients. Discrimination is a different treatment of individuals (patients), based on morally irrelevant criteria.

The idea that moral imperatives must be equally binding on everyone is an example of universalization with respect to the agent. Kant’s famous categorical imperative (Kant, 1785, p.30), can just as well be understood as a universalization with respect to agents and/or patients. In its ‘universal law’ formulation, the categorical imperative goes as follows: “act according to that maxim (moral rule or guiding principle⁴) whereby you can will that it should become a universal law.” One should ask the question: “What if everyone (or many people) acted (or thought it is allowed to act) in this way?” There might be two contradictions: a contradiction in conception, which means that the universalized rule results in a logical or physical impossibility, and a contradiction in

⁴ A note on terminology: most of the time, a ‘principle’ refers to a ‘basic principle’. This dissertation argues that there are five basic principles. The first basic principle is called ‘rule universalism’. This principle refers to the universalization of moral ‘rules’ or ‘guiding principles’. Those moral rules are typically less basic, less abstract or more specific than the basic principles. Moral rules are derivatives of basic principles. For example the moral rule “Do not steal” is derived from the basic principle to maximize aggregated well-being. The moral rule “Do not rape” is derived from the basic mere means principle.

the will, which result a universalized rule is possible, but not wanted by a rational moral agent.

An example of a contradiction in conception, resulting from universalizing a rule with respect to the agent, is the ecological footprint. If everyone (all humans) would consume as much as an average human in a developed country, we will exceed the carrying capacity of the planet. We would need resources of more than one planet, because the ecological footprint in the developed world is higher than the available biocapacity of 1,8 global hectares per human (GFN, 2010). There is no other planet like Earth, so this behavior is simply not universalizable. Therefore, it is unfair (and unsustainable) for those people in developed countries to have such a high ecological footprint.

A contradiction in the will might occur after a universalization with respect to both agent and patient. For example if I am allowed to harm you, then everyone is allowed to harm everyone else. So you are then also allowed to harm me. I cannot will this, because I value my well-being, so I encounter a contradiction. My rule that I am allowed to harm you cannot be universalized, so it is an immoral rule. This universalization therefore generates the golden rule: (do not) treat others as you would (not) like to be treated.⁵

Universalization with respect to the act becomes much more subtle. Note that the above expression (“You must follow the rule...”) refers to rules instead of particular acts⁶. Hence we can call it rule universalism. A formulation in terms of rules instead of acts has some advantages. First, it allows for conditional rules (e.g. “Do X if Y unless Z”)⁷. If I say that I am allowed to lie down on my sofa and watch a movie tonight, this act cannot be universalized with respect to agents, because it is impossible for 7 billion people to fit on my sofa tonight. We might save the idea behind the categorical imperative if we universalized a rule instead of an act, because the rule can contain some conditionals. I can follow a rule which says that I am allowed to sit in a place I prefer and watch a movie at a time I prefer, if I own the place (e.g. my sofa) or if there is some place left (e.g. in a movie theater). Universalizing this means: everyone is allowed

⁵ The following example is a critique to an oversimplified application of the golden rule. Suppose that I don’t mind if you lied to me. One might interpret the golden rule as claiming that in that case it is allowed for me to lie to you. However, this application of the golden rule is itself a violation of the golden rule. We can derive a ‘platina rule’ from the golden rule: do not take your own preferences when you decide how to treat others, because you would not like others to take their own preferences when they decide how to treat you. In other words: I am not allowed to lie to you if you do not want be lied to, because I would not like you to lie to me when I do not want to be lied to.

⁶ We encounter a similar difference between rules and acts in the discussion between rule consequentialism and act consequentialism (Hooker, 2011).

⁷ And it allows for game theoretic situations where the choice what to do depends on what others do or should do (see section 7.2).

to lie down in a place they prefer and watch a movie they prefer, if some of those conditions are met. A focus on rules instead of acts allows a specification of the conditions that can be universalized in such a way that we can want this universalization of the conditional rule.

A second advantage is that the focus on rules allows for public expressions, such that all moral agents are able to understand and follow the rules. Rules are useful tools in giving justifications of acts. As a consequence, a focus on rules gives meaning to the idea of giving the good example as well as the idea of giving the right justification.

Universalization also applies to metarules (rules about rules). For example, I can propose the universalized rule: “Everyone should always tell the truth, unless your name is Stijn Bruers.” But now Marie can use a universalization on the level of metarules and respond that if I am allowed to use my name in a rule, then she is allowed to do the same: “Everyone should always tell the truth, unless your name is Marie.” I do not want this, and as a consequence, we can derive a metarule: “Rules should not refer to names.” Similarly, rules should not refer to nouns or specific times and places. This universalization on the level of metarules gives important clues on what counts as morally (dis)similar situations. A situation where Stijn lies is similar to a situation where Marie lies.

The universalization criterion matches two conditions: non-arbitrariness and consistency. An ethical system based on rules that are not universalized may easily be consistent, but the lack of universalization results in arbitrariness. Universalizing the rules avoids the arbitrariness, but now the system might contain mutually inconsistent rules. Both non-arbitrariness and consistency should be respected. A rule should be non-arbitrary (i.e. universalized) and consistent with other non-arbitrary rules of the system. The two conditions of non-arbitrariness and consistency generate coherence (see Chapter 2).

Universalization in ethics is related to the property of supervenience of moral statements. Supervenience is a kind of dependency relation: moral judgments supervene on natural properties in the sense that if two situations, acts or events are similar in their natural properties⁸, then they should imply the same moral judgment (or moral value). Or stated in reverse: if you have different moral judgments about two different situations, then you should be able to point at a morally relevant, objective difference between those two situations. Compare this with the supervenience of mental states on physical brain states: psychological properties supervene on physical properties in the sense that *all* persons who have the same physical properties (the

⁸ With “similar natural properties” I mean similarity only in their relevant aspects.

same brain states) must also be psychologically indistinguishable (having the same mental states).⁹ Note the universality of this statement.

Note that in non-cognitivist meta-ethics such as expressivism, supervenience is considered as a consistency condition: expressivists claim that moral judgments are based on our subjective attitudes toward objective, natural properties (e.g. behaviors or situations). Supervenience means that our attitudes should be consistent and non-arbitrary.

The process of universalization is omnipresent in ethical discussions and argumentations, because universalization allows for the use of analogies, and analogies are often used in argumentations. People frequently refer to another specific situation which is similar to the discussed situation, and then appeal on intuitions in that other situation. The question is: when is an analogy valid?

It is clear that the validity of the use of an analogy needs to be justified as well. But first, of course, we need to justify the use of the method of universalization. To do this, I will refer again to the six analogies of ethics mentioned in the previous section. In all those examples, universalization is important. Next, I will demonstrate that a lot of people, including meat eaters, use this method in ethical discussions. I will present a list of arguments given by meat eaters, in order to demonstrate that meat eaters also accept this method of universalization. Hence, after accepting the importance of moral intuitions, we come to see that animal rights ethicists and meat eaters both share the same rules of the game. This should bring a compelling rational argument for veganism and animal equality one step closer.

So let's first take another look at the six analogies.

1) Physics and other empirical sciences. We had a specific observation that at time T at place P under conditions C, I saw object O falling. I now make a hypothesis which goes under the name of induction: *all* similar objects fall under similar conditions at *all* times in *all* places. So when I see another object falling when I release it, it is conforming to this principle of gravity. If it is not falling, then something relevant must be different: the place (e.g. in a space ship), the conditions (there is a strong force such as a magnetic field counteracting gravity) or the object (it has a propeller and wings, or it is a balloon with helium). When the object does not fall, it simply means that other principles should be included and those principles need to be universalized as well (e.g. *every time* an object is released in the presence of a strong magnetic field that acts as a force...).

⁹ A difference between psychological supervenience and moral supervenience is that the latter has no physically or logically necessary relationship. In this sense, supervening moral judgments are not facts of the world, whereas psychological states are facts of the world.

2) Mathematics and algebra. From the intuition that the number 3 has a successor, we universalize this to the mathematical axiom that *every* number has a successor. This is a basic axiom in the algebraic system of natural numbers. Mathematicians do not allow an exception to this rule. But there can be other universalized rules with exceptions, and of course those exceptions need also to be universalized.

3) Geometry. *All* lines of equal length are similar. *All* right angles are congruent. (The latter principle is one of the basic Euclidean axioms.)

4) Grammar. All sentences with the structure “Subject + verb + predicate” are grammatically correct. “The gnorfl is sprinkle” was just one example of such a structure. But there is an exception: the structure is not correct when the sentence is a question. And this exception rule needs also to be universalized: the structure “Verb + subject + predicate?” is always correct in case of a question.

5) Crossword puzzles. If the input data is “fruit” and “five letters”, and we can fill in the letters “APP”, then we are forced to complete this word to “APPLE”. So filling in a word gives information about the content of *all* relevant white boxes, i.e. all boxes that are similar. Similarity means that they are next to each other in a row or column. This is also a kind of universalization. An individual letter represents a particular ethical rule, applicable to a particular situation. A word is analogous to the universalized principle, applicable to all similar situations.

6) Taste evaluations and esthetic judgments. This pear tastes good to me, so now I have to accept that *all* products with the same chemical structure taste as good. But there is an exception: e.g. when it is hot, it no longer tastes good. So we again universalize this to the principle that all products with a similar chemical structure and the same temperature taste as good.

In summary: universalization is perhaps the most important tool in moral arguments, because it is the basis of consistency. Universalization happens not only in ethics, but in all other cognitive activities, as long as similar conditions apply (this is also a universalized statement). If animal rights ethicists want to convince meat eaters that we should become vegans, and if these ethicists want to use rational arguments, meat eaters should accept the same rules of the game. Meat eaters already accepted moral intuitions as starting points, so let us see whether they can also accept the method of universalization. Looking at discussions with meat eaters, it becomes clear that those ingredients are indeed important for them as well. Both meat eaters and animal rights activists use all four forms of universalization in their arguments. Here is a list of arguments by meat eaters. They indicate that meat eaters value universality and consistency as well.

1.2.1 Universalizations made by meat eaters

1.2.1.1 Universalization with respect to the situation or the act

“But aren’t you wearing leather shoes?” If killing animals for food is not allowed, then killing them for clothing is not allowed either. Or if you are allowed to wear leather shoes, then you are also allowed to eat meat. This is a correct argument, and it is the reason why ethical vegans won’t buy leather shoes.

“You are a vegan, so you are also against breast feeding?” If you are against the use of animal products, you should be against breast milk as well, because humans are animals as well. This is an invalid argument, because we can easily point at morally relevant differences: breast milk is necessary for the baby, and the mother gives it voluntarily. The ethical rule “Do not use animal products” has exceptions, such as necessity and voluntariness. These exceptions should be universalized as well.

“So we shouldn’t walk around because that will kill insects?” Killing cows for food is like killing ants by walking around. This is an invalid argument, because of two reasons: insects likely lack or have a very small emotional life, and killing by accident when walking around is different from intentionally killing someone in order to use him.

“You are against animal experiments, so you never use drugs tested on animals?” This is an invalid argument, because of two reasons: 1) animal rights activists are only concerned about animal experiments done in the present and future, and 2) meat eaters are likely to use products whose development or discovery involved rights violations in the past that they abhor. For example they might use technology based on mechanisms or materials that were discovered with the help of slaves some hundred years ago. Note that I hereby replied with an analogy, referring to technological inventions and slavery. We can argue that the analogy is valid, as the relevant properties are similar: rights violations of animals and slaves, the use of animals and slaves in the process of developing something that we now use. So I replied with a universalization.

“Animal rights activists want to promote freedom of animals, but they restrict our freedom by imposing veganism against our will. That is inconsistent.” This is an invalid argument because of two reasons: 1) there is a morally relevant difference between killing someone in order to enjoy the taste of eating him, and restricting someone’s freedom to violate rights. 2) Meat eaters themselves promote freedom of e.g. women by restricting the freedom of rapists and imposing their ethics upon them. Again I make an analogy, which is customary in a process of universalization. I universalized with respect to the patient, from animals to women, and with respect to the act, from slaughter to rape. The analogy is valid, because raping women and killing animals are both examples of violations of bodily integrity just for pleasure, and these are morally relevant facts.

“Animal rights activists use the earth as well, kill life, rob land from wildlife, eat food that otherwise animals could have had... The production of vegan food also kills animals.” This argument is invalid because there is a morally relevant difference between killing by accident and intentionally killing someone. Meat eaters also eat food that otherwise other humans could have eaten. And animal rights activists consistently claim that we do have a duty to protect wildlife animals as much as possible, and help them if we can.

“Rights are a human invention, so talking about animal rights is still anthropocentric. Therefore it is inconsistent with the idea that anthropocentrism is bad.” This argument is invalid, because this is a confusion of two different notions of anthropocentrism, and these two notions are morally not the same. If rights were my invention, this is egocentric, because rights originated from me. But that kind of egocentrism is not a moral problem. However, it is different from the dangerous egocentrism that claims that I am the only person in the world who has rights.

“If we give animals the right to live, then we should also give them the right to vote.” This is an invalid argument because of two reasons: 1) higher intelligence is a capacity that is clearly relevant in the right to vote but not the right to live, and 2) meat eaters themselves give babies and seriously mentally disabled persons a right to live but not a right to vote.

“If animals have rights like humans, then we should also need a huge number of animal hospitals. That is unrealistic.” This argument is valid to some degree: just like helping humans we have a duty to help wild animals in need, as far as this is feasible for us. There are already wildlife rescue centers (I happen to do voluntary work in a bird care center), so it is not unrealistic.

“Animals mate with conspecifics. So do humans. Isn’t that speciesism as well?” I leave it up to the reader to find the morally relevant difference in this case.

“What’s wrong with the pleasure of the taste of meat? Should we forswear all pleasure?” This argument contains a universalization: if A is bad and done for pleasure, then anything done for pleasure is bad. However, that rule quickly violates intuitions that both meat eaters and vegans share. So the animal rights ethicists come up with a more accurate universalization that says that we should forswear all pleasure if important rights are violated.

1.2.1.2 Universalization with respect to the patient

“What about plants, insects,...? They also have a life.” The morally relevant difference is sentience: plants and insects have a much lower probability to be sentient than for example vertebrate animals with a complex functioning central nervous system. In a later chapter (section 8.5) I will give several arguments why sentience is morally relevant in this matter.

“Animals can’t think rationally, have no self-consciousness, have no moral agency,...”

Here the meat eater tries to point at some special mental capacities that animals lack. However, also some humans lack them, but most meat eaters give basic rights to e.g. mentally disabled persons.

“Eating plants is allowed because it is natural and normal. Eating meat is natural and normal as well, therefore it is also allowed.” Here the meat eater uses a rule that applies to plants, and extends it to animals. The problem with this argument is that the criteria ‘normal’ and ‘natural’ are vague. Depending on how one might more accurately define those terms, one can argue that eating humans would be allowed, or rape would be allowed. More about the two criteria normal and natural will be discussed in the chapter on predation in part 3.

1.2.1.3 Universalization with respect to the agent

“If I am not allowed to eat meat, then lions, the Inuit people,... are not allowed to hunt either?” Here the meat eaters says that if A (a lion) is allowed to eat meat, then B (a human in a developed country with a moderate climate) is also allowed to eat meat. In the chapter on predation in part 3 I will elaborate on this argument, and show that the universalization is not valid, because an important necessity criterion is not universalized.

“If everyone (all humans) would be vegan, then it is impossible to feed everyone. Agriculture without livestock is impossible (for example no manure to fertilize the croplands,...)” This argument points at a Kantian contradiction in conception. It is actually a factual claim. There are studies however that indicate that a global vegan organic agriculture, without use of chemical and animal fertilizers, is feasible and can feed a world of 9 billion people (Fairlie 2007; Olewski, 2010; the Vegan Organic Network). Typically, vegan products with the same nutritional value require much less inputs such as water, energy, land and chemicals than livestock products (GFN, 2010; Hoekstra, 2010). And they generate much less negative outputs such as nitrogen pollution and greenhouse gas emissions.

“If everyone (all humans) would eat vegan, then all animal races that we breed in the livestock sector would go extinct. That’s also a loss of biodiversity.” The problem is that livestock currently is likely the largest threat to wildlife biodiversity (FAO, 2006). And secondly, is it really a contribution to biodiversity to intentionally breed animals with serious physical handicaps? That is not morally justifiable.

“Our ancestors ate meat. It would be unfair towards us if we would not be allowed to eat meat.” This is a universalization from our ancestors to us. My counterargument is a universalization with respect to the act. If such an argument would be valid, then a similar argument would be valid for slavery, rape,... We would be allowed to do that

because likely one of our ancestors did it. We cannot will this universalization, and therefore we have to say that what our ancestors did cannot justify what we do.

“If everyone (all humans) would be vegan, all people working in the animal food production would lose their jobs.” However, new jobs would be created to produce vegan food. In the end it comes down to the fact that vegan farming is much more efficient in terms of inputs, and therefore less capital intensive. In other words: vegan farming makes much more sense economically. And a second counterargument, based on the universalization of the act: a same thing has been said about slavery and e.g. slave traders losing their jobs. If slavery cannot be justified by pointing at (invalid) economic concerns, then animal food production cannot be justified either.

1.2.2 Universalizations made by animal ethicists

Of course, in discussions, we often hear animal ethicists giving arguments based on universalization as well.

1.2.2.1 Universalization with respect to the situation or the act

“Speciesism is similar to racism and sexism.”

“The livestock industry is similar to slavery, rape,...”

“If we are allowed to eat cows because we breed them for that purpose, then slavery would also be allowed if we breed black people for that purpose.”

“If we are allowed to eat cows because they owe their lives to us, then we are also allowed to eat our babies or humans that we would breed.”

“If we are allowed to use animals in the livestock industry because they have food, shelter and medicines and hence those animals are better off than animals in the wild, then it is also allowed to use indigenous people as slaves, because in their natural habitat they suffer from predators, parasites, drought, disease,...”

“If mentally disabled persons get rights because they belong to a group whose normal members have rationality, then also a mentally disabled person should have a right to vote and go to university. And then also chimpanzees and other primates should have rights, because the majority of them (roughly 7 billion primates) has rationality.”

“If meat consumption is permissible because it is not prohibited by the law, then also slavery, rape,... were permissible in times when it was not prohibited by law.”

“If eating meat is allowed because it is natural, then rape should also be allowed, because it is also natural: our ancestors did it for thousands of years, other animals do it, men have developed (by evolution) special body parts that enable them to rape women, and it is very likely that we owe our lives to the fact that one of our ancestors once raped a woman...”

“If veganism is not good because it is unnatural in the sense that one needs vitamin B12 as supplements (and those B12 products come from a factory), then eating processed foods or using toothpaste would also be unnatural and bad. The use of toothpaste is a consequence of our choice of diet, just like the use of B12 is a consequence of a dietary choice. And both are produced in a factory. It’s even worse for the non-vegan person, because drinking cow milk at an adult age, animal experimentation, feeding B12 supplements to livestock animals, and much more would also be unnatural and hence bad.”

1.2.2.2 Universalization with respect to the patient

“If eating pigs is allowed, then eating dogs should also be allowed.”

“If killing and eating animals is allowed, then killing and eating humans should also be allowed.”

“If animals don’t have rights because they don’t have duties and they cannot understand ethics, then the mentally disabled persons also cannot have rights.”

1.2.2.3 Universalization with respect to the agent

“If you don’t need to harm sentient beings in order to survive, then you should not harm them. Of course that also applies to me, so I became vegan.”

“If I am the only vegan person, the market will not notice my food choice, and not a single cow will be spared. However, I should give the good example and do what everyone should have to do. If everyone became vegan, then animal rights violations would be much lower and that is good.”

The above examples show that the method of universalization sets a common ground for both animal ethicists and meat eaters. But are people always consistent in applying the technique of universalization? Do they sometimes use this technique and other times not? My claim – based on what we have seen a little bit in the above examples – is that especially meat eaters (and sometimes also animal ethicists) do not consistently use the technique of universalization. Only when applying the technique of universalization consistently (universalize the technique itself, so to speak) is it possible to reach a coherent ethical theory.

Chapter 2 The goal: consistency and coherence

In the above section, we started by looking for moral intuitions. That was the non-reflective part. At the reflective level, these intuitions were first articulated as particular ethical rules, and second, these particular rules were universalized to all similar situations (acts, patients and agents). Hence, we moved from moral intuitions to particular ethical rules to universalized ethical principles. Now we have to see whether those universalized ethical principles are internally consistent. If there is an inconsistency, we have three options: we can either introduce a new principle that overrules the other, we can refine an existing principle or we can simply delete the weakest principle (which is based on the weakest, least motivating moral intuition).

By testing more and more situations, i.e. by looking at whether our moral intuitions in all situations agree with the universalized ethical principles, we progressively get a coherent system. This is the process of reflective equilibrium (Rawls, 1971). Coherence means that several intuitions and principles enforce each other; an interweaving of mutually supporting intuitions and principles.

Here the analogy between constructing an ethical system and solving a crossword puzzle becomes clear. In the construction of an ethical system, we start with our strongest moral intuitions; those intuitions in which we have the most confidence. Similarly, in the crossword puzzle, we have more confidence in a word if there are less other words that fit the description. If for example the description is “food” of five letters, then the word can be “APPLE” or many other possibilities. If however the description is reduced to “pomaceous fruit that grows on a tree”, we are much more confident that the word should be “APPLE”. Once we fill in some words, we get new evidence for other words. Suppose I have filled in the word APPLE. The first letter crosses another word of five letters, with description: “good”, and with the new information that the fourth letter should be an A. So I can fill in the word MORAL. The last letter of this word crosses a second word of two letters, a music note: LA. And now the P from APPLE and the A from LA gives us a new clue about a body organ with eight letters: PANCREAS. The whole set-up is not only consistent, but those four different words enforce each other. We gain more and more confidence in the whole system.

Coherence is the combination of three things: *universalism* (or *non-arbitrariness*), *consistency* and *completeness*. Universalism in the crossword puzzle means that all

neighboring white boxes in a row or column should form one word; not just independent letters or smaller words. If the white boxes were filled in with independent letters instead of meaningful words, the crossword puzzle would become completely arbitrary. The constraint that the rows and columns of white boxes should be words, decreases the arbitrariness. The same goes for ethics: similar situations (or the same situation from different points of view) should be judged with the same ethical principles. If every situation or viewpoint has its own particular guiding principle or moral rule, the ethical system becomes too arbitrary. Universalism strongly decreases arbitrariness.

Consistency in the crossword puzzle means that a white box cannot contain more than one unique letter. In ethics, it means that each situation should have one solution, one “all things considered” moral judgment. If in a situation there are two different inconsistent solutions (two different judgments about e.g. the (im)permissibility of action), the ethical system is inconsistent and we are left undecided. In the crossword puzzle, there is a difference between trivial consistency and strong consistency. Trivial consistency occurs when words do not cross and there is only one cue for a word. For example if the given cue is “fruit” and the word contains five letters, “apple” is a trivially consistent solution. Strong consistency occurs when words cross each other and there are more cues, e.g. for vertical and horizontal words. If the first letter of the word “apple” is crossed by a second, vertical word of six letters with cue “nut”, then the letter “a” in “almond” is strongly consistent with the same letter in “apple”, because the letter “a” has two justifications. In this dissertation, consistency always means strong consistency. For example, as we will see, when an ethical principle of well-being has two different justifications derived from two different points of view, one based on a thought experiment of impartiality and one based on the virtue of compassion (see section 4.4), this principle becomes strongly consistent.

Completeness, as the third condition, means that in the crossword puzzle no white box should be left empty. In ethics, it means that the ethical system should be applicable to all possible situations and viewpoints.

This idea of coherence and crossword puzzles was proposed by Haack (1993). She called it ‘foundherentism’, as it is a combination of foundationalism and coherentism. The foundations are the input data: information about the meaning and lengths of words. Compare this with our moral intuitions or with mathematical axioms that also act as foundations. But not all our moral intuitions are equally strong. Some are much more motivating, others are easily overruled. Most of all: all our intuitions are fallible. Our moral intuitions are merely provisional starting points in the construction of a coherent system. They are still subject to revision or rejection if they are not compatible with the rest of the ethical system. In the construction of a coherent system, there is no infinitely strong moral intuition that can serve as the foundation. There is no absolute fixed point. Similarly, in a crossword puzzle we can have different levels of confidence

in the words, and there is no central word from which to start. In principle, all solved words are fallible; we can never be sure that a word in the puzzle is absolutely correct. To gain confidence, different words can mutually support each other, and build up a coherent system. This is the coherentist part of the story. Both foundations and coherence are important, but can be important to different degrees.¹

This principle of coherence (moving towards a reflective equilibrium with principles that mutually support each other) is also universal, as it occurs in our six analogies as well:

1) Physics and other empirical sciences. Evidence builds up. A theory is a coherent system. And sometimes new strong data from experiments appear that is really incompatible with the principles. So the theory needs a revision. And like a crossword puzzle, sometimes other aspects of the theory need revision as well. It can ignite a cascade of revisions. Even the principles behind the reliability of the experimental apparatus itself might need revision. This is characteristic of scientific revolutions (Kuhn, 1962). But not to worry, as the more coherent a theory becomes, the less likely we need to revise the whole thing. The more words are filled in the crossword puzzle, the more likely it is the real solution. The scientific method is nothing but a constant searching for a reflective equilibrium: a coherent, parsimonious, clear system of knowledge that best fits the most convincing input data. Anomalous data that do not fit a strong coherent scientific theory can be discarded, as they are likely the result of some error. (See Brophy, 2009, for an elaboration on the analogy between the scientific method and the moral method of reflective equilibrium).

2) Mathematics and algebra. Mathematicians often wondered whether the system of natural numbers is a consistent system. Gödel (1931) showed that its consistency was impossible to prove from within the system. One needs to extend the system (step outside of the system) in order to prove the consistency of the system of natural numbers. But then the question reappears: is this extended system consistent? Anyway, we do notice that the system of natural numbers is strongly coherent. Let's take the example of the property that the sum of integers from 1 to N equals $N(N+1)/2$. You can test this several times, taking $N=3$; $N=10$; $N=17$ and so on. The more you try it, the more plausible it seems. These tests are coherent with a proof that one could give: a proof of induction. The property is true for $N=1$ and $N=2$. Suppose it is true for some arbitrary N . Then for the number $N+1$ we get the sum from 1 to $N+1$ should be $(N+1)(N+2)/2$. And

¹ During the construction of an ethical system, no intuition or principle is absolutely fixed. During the process of solving the crossword puzzle, no letter or word is absolutely fixed. However after the crossword puzzle is completely solved, we can be (almost) absolutely confident in the completed words. Similarly, after we have constructed a coherent system, the final universalized ethical principles can be considered as foundations or basic principles.

indeed, this equals $N(N+1)/2+N+1$. This proves the proposition. And then one could give a second proof and a third. So these different observations and proofs indicate that the system is strongly coherent, which increases our confidence in its consistency.

3) Geometry. One could test the Pythagorean theorem with some examples of right-angled triangles. And then one could give different proofs of it (there exist at least 370 different proofs of this famous theorem (Loomis,1968)).

4) Grammar. Grammatical rules might give a coherent picture of our intuitions as well, by analyzing long sentences, breaking them down in parts according to a grammatical rule.

5) Crossword puzzles. This example I already explained above.

6) Taste evaluations and esthetic judgments. The fact that I like pears is consistent with the fact that I like other fruits, the fact that I liked a particular pear yesterday, the knowledge that pears contain sugar and I also like sugar, and so on. So this gives me a coherent picture of what I like.

In ethics it is possible that there exist mutually incompatible ethical systems, all internally consistent and based on some moral intuitions. This might point at some kind of ethical relativism, because each consistent system is equally valid. However, let's go back to the analogy of the crossword puzzle. Here it might be possible that a crossword puzzle has multiple solutions: different mutually incompatible patterns of words. But the more words a puzzle contains, and the lengthier the words, the less likely that there are different solutions. I believe that if we take the set of strongest intuitions that both I and a meat eater share, then the only coherent ethical system that we can construct, is one that implies veganism. Of course there are meat eaters who have totally different moral intuitions. I will never be able to convince the latter by merely rational arguments, because we start from different input data. So I want to address myself to those meat eaters who have some strong moral intuitions shared with me. And looking at discussions with meat eaters, I notice that they also value consistency. Why else would they ask me if I'm wearing leather shoes? Why else do they so often come up with arguments based on universalized principles, as we've seen in the previous section? Consistency is important for a lot of meat eaters. So we can agree on the goal, the method and the input. We agree on all elements that might settle the issue of the permissibility of eating meat.

Of course, consistency is not the only goal. Suppose someone claims that s/he feels a strong intuition that gay marriage is immoral because it is impure. So s/he adds this rule to the ethical system, as a dominating principle (it dominates the principle of well-being of gay people). The resulting theory is now consistent. Yet, the problem is that this criterion of purity is not clear at all. I cannot understand this principle (how can I detect when something is impure?), thus I am not able to universalize the principle and test it in other situations. Therefore, a principle based on an intuition should be very

clear, so that people who do not have that intuition are still able to understand it and test it in different situations and thought experiments. In the second part of this dissertation, I will present some principles, such as basic rights, which are based on intuitions. As we will see, I elaborate on criteria to test whether a basic right is violated. The basic right principle is and should be formulated as accurately as possible, so that even a computer might be able to test it. This precision allows the principle to be tested in well-constructed moral dilemmas and thought experiments.

In the end, the anti-gay person is requested to formulate his purity principle as clear as possible, so that we are able to universalize it and test the consistency of this theory. Personally, I doubt whether one can make the system consistent by adding such a purity principle.

To conclude, what I want is a coherent ethical system, consisting of clear and mutually consistent universalizable principles that best fit our shared, strongest moral intuitions, without adding too many arbitrary elements. That is the goal: a theory in coherent reflective equilibrium. It really reflects the scientific approach. In science, the hypotheses should be formulated as clearly as possible, in order to test them in experiments. The theory should be as parsimonious as possible, and of course internally consistent and consistent with the most reliable observations and test results.

There is one big problem however, a problem we seriously have to deal with: illusions.

Chapter 3 The problem: moral illusions¹

The question is whether our input data (moral intuitions) can be trusted. Are they always reliable? Unfortunately, they are not. As in science, we have to admit that some of our observations and experiments were unreliable. In search for a coherent reflective equilibrium, we might come to the conclusion that some of our input data were wrong, meaning they cannot be reconciled with the rest of the constructed system. After stumbling upon a contradiction, there are two options. First our intuitive moral judgment might easily change or disappear. Such judgments will be called moral mistakes or deceptions. In my personal experience, the judgment that it was allowed to eat meat, was a moral deception. I now have developed new insights, a new coherent system, and it happened that my previous judgment that we are allowed to eat meat, was incompatible with this ethical system. That old judgment completely disappeared; I changed my mind.²

But there is another possibility: sometimes our intuitive judgments do not seem to disappear, although we have to admit that they are not compatible with a strong coherent system. These moral intuitions will be called moral illusions. They are analog to optical illusions. Those optical illusions are characterized by their so called cognitive impenetrability (Pylyshyn, 1999). Even after I know that my visual perception of a figure is wrong, the illusion persists. Similarly, even after I realize that a moral intuition cannot be included in a strong coherent ethical system, that moral intuition might still

¹ Sections 3.1, 3.2, 3.3 and 3.4 are based on *Moral Illusions in the Trolley Problem*, Bruers & Braeckman, submitted to Croatian Journal of Philosophy.

² In the method of reflective equilibrium, one first starts with a filtration process: eliminating the initial judgments (moral intuitions) which are clearly unreliable, such as the intuitions that arise in conditions disposed to error (e.g. heavy emotional influences, morally irrelevant situational elements that trigger feelings of disgust,...). What is left is the set of considered (credible) moral judgments, and this set of credible moral intuitions is used in the construction of a coherent system in reflective equilibrium. See e.g. Brophy (2009), who compares this filtration process with the scientific practice of data selection, rejecting those data that were gathered in error-disposed conditions. My hypothesis is that the rejected moral intuitions (those that lack some initial credibility) are examples of moral mistakes or deceptions. The set of considered moral judgments might still contain moral illusions.

persist and ‘stay alive’. If that incompatible moral intuition is weaker than the coherent ethical system (if we are less willing to give up the whole ethical system), we should declare this outlying intuition to be a moral illusion.

As optical illusions exist, moral illusions might also exist, and they might have cognitive impenetrability just as well. In fact, I will argue in a later chapter that speciesism (the intuitive, prejudicial judgment that humans have a higher moral status than non-humans) is a moral illusion. If speciesism is a kind of discrimination but is cognitively impenetrable, this explains why it sometimes is so difficult to convince a meat eater that veganism is a moral duty.

Moral mistakes are rather easy to deal with, because they will simply disappear. But the cognitive impenetrability of moral illusions is a tougher nut to crack. We can’t just take all of our intuitions for granted. But how do we know whether an intuition is an illusion or not? To clarify this, let’s first look at how we deal with optical illusions.

3.1 Optical illusions

A paradigmatic example of an optical illusion is the Müller-Lyer illusion (Müller-Lyer, 1889). This figure consists of two horizontal line segments, with inward and outward pointing arrowheads (see figure). The line segment with the outward pointing arrowheads appears smaller than the other line segment. In other words: our intuition judges the lower horizontal line to be longer than the upper one.

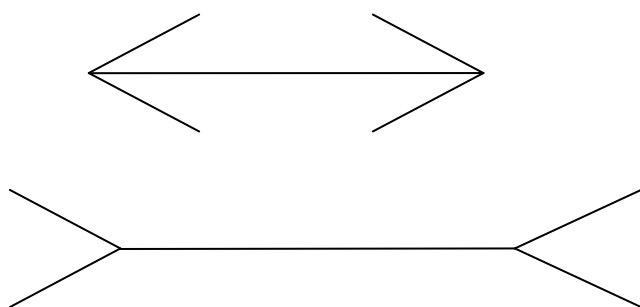


Figure 1: the Müller-Lyer optical illusion

But how do we actually know that this intuition is illusory? Of course we could simply refer to an objective reality to claim without further justification that both lengths are equal. The lengths in the Müller-Lyer figure are primary qualities, i.e. factual, objective properties of the world. A correspondence theory of truth states that our judgment about the lengths is either correct or incorrect, depending on whether the lengths in

the figure are objectively equal or not. But that would presuppose that we could get direct knowledge of this objective reality. Furthermore, I want to study moral illusions, and in ethics there are moral non-cognitivists who do not believe that moral judgments state true or false facts about an objective moral realm.³ In that case, a simple correspondence theory of truth may not be applicable when it comes to moral judgments. It is better not to refer to objective realities, because the strategy to find moral illusions should be satisfying for non-cognitivists as well. We should be able to study moral illusions even when a realm of non-subjective moral facts and truths does not exist. The approach I suggest is based on coherentism that uses intuitions instead of objective facts as input data to construct a coherent system. A coherence theory of truth, instead of a correspondence theory, should do the job.

To prove the illusion in the Müller-Lyer figure, we first have to make all our most evident intuitions about geometry explicit. Then we will see that this one intuitive judgment about the lengths in the Müller-Lyer illusion is in contradiction with two other, stronger intuitions.

The first of those stronger intuitions says that this ruler does not change its length when it is shifted in this direction. This translation invariance intuition seems obvious, but it is impossible to give a further argument to prove this⁴, so therefore it is a basic optical (or geometrical) intuition. This intuition is expressed and universalized into the very important geometrical principle of translation invariance: all rulers keep their length when shifted in any direction. This is a universalized principle that is true for all rulers and all translation directions in all situations. So it should apply to the above figure as well. We accept this principle as self-evident, although it is possible to think of hypothetical worlds or complex geometrical systems where translation invariance is not valid (mathematicians already constructed lots of counter-intuitive geometrical systems: e.g. projective, non-Euclidian or non-commutative geometries).

A second intuition says that the length of this line segment does not depend on the presence of these other lines around. This intuition is translated into the very important universalized principle of context independence: all line segments have lengths independent from any geometric objects around. This, too, is a universalized principle.

³ I don't want to defend an intuitionist or naturalist meta-ethical position. These positions are realist-cognitivist, in the sense that they claim that there exist moral facts in the world, that we can get access to (or knowledge of) this realm of moral facts and that some of these facts are true. The approach that I defend is constructivist: we construct coherent ethical systems. I leave in the middle whether the constructed coherent system is a representation of some objectively existing ethical system in the world.

⁴ With "further argument" I mean an argument based on another foundational principle or intuition. Of course, the translation invariance intuition in a particular situation is coherent with similar intuitions in other particular situations, but this coherence is not what I mean with a "further argument" for its validity.

The arrowheads are the ‘context’, and the claim is that lengths are *always* independent from *any* context (any other thing floating around).

This principle of context independence is also related to an intuitive aversion for *arbitrariness* and *artificiality*.

Arbitrariness has two aspects: a vertical and a horizontal one. *Vertical arbitrariness* says that it is arbitrary to claim that lengths of line segments are influenced in ‘four-legged’ figures (two arrowheads with two legs each) instead of ‘N-legged’ figures or figures with other objects than arrowheads (e.g. line segments with circles instead of arrowheads). *Horizontal arbitrariness* says that it is arbitrary that, within the class of four-legged figures, the length of a line segment decreases when the arrowheads are pointed outwards instead of inwards. Compare this double arbitrariness with a wardrobe containing vertically arranged drawers. Vertical arbitrariness claims it is arbitrary to select the third drawer with pants. Horizontal arbitrariness says that, in this third drawer, it is arbitrary to select the brown pants instead of the blue.

In section 8.4, I will demonstrate the vertical and horizontal arbitrariness of speciesism. Also other examples can be given, such as in religious and creationist beliefs. The aversion for arbitrariness is a strong motivation for atheists. Consider a book: it is unlikely that the book is spontaneously written, because it has a high information content. So it is created by a rational being, an author. Creationists say that this author is also too complex to have been spontaneously generated (and those creationists do not believe in the third option next to spontaneous generation and conscious creation: natural evolution). So they believe that there should exist a meta-author, a god who created the author (and the rest of the universe as we know it). The problem is that most creationists stop at this second level of creation: they say that what exists is ‘universe+god’. But of course, by the same reason ‘universe+god’ is also too complex for a spontaneous generation, so it should have been created by a metagod. But ‘universe+god+metagod’ is again too complex, so there should be an even higher creator. We end up with an infinite chain, a recursive set of sets of creators. Most religious believers pick the second lowest level, where a god created creators such as authors of books. This choice for the second level is arbitrary, because the evidence for this level is not stronger than the evidence for any higher level (only the evidence for the lowest level, the level of authors, is very strong). This is the vertical arbitrariness, which reflects the famous creationist problem: who created god?

Within this second level, there is a further horizontal arbitrariness. For example a Christian creationist does not believe in the many other possible creators and gods: s/he does not believe in Brahma, Thor, Jupiter,... As the evidence for a Christian god is as high as the evidence for e.g. a Hindu god, the choice for the Christian god is arbitrary. The Christian believer is in fact atheist about all those other possible gods. The arguments raised by a Christian believer towards atheists bounce back: “You are not open for God” (neither is s/he for Krishna), “You cannot prove that God does not exist” (neither can

s/he prove that Quetzalcoatl does not exist), “God reveals Himself once you believe in Him” (but so will Osiris), “It’s arrogant to claim that God does not exist” (but also arrogant to claim that Zeus does not exist).

Artificiality or *complicacy* claims that the influence of e.g. the angles of the arrowheads on the length of a line segment generates a geometrical rule that is too complicated and farfetched. This artificiality introduces a fuzzy factor: what if we gradually open the angles of the arrowheads? How should this influence the lengths of the horizontal line segments? Such mysterious influence seems artificial.

In section 8.4, I will argue that speciesism is very artificial. Again we can make an analogy with e.g. religious beliefs: looking at religious doctrines from an outsider’s perspective, they often look highly complicated and farfetched. Atheists can ask lots of puzzling questions about religious doctrines, such as: why did God put the forbidden tree of knowledge in the middle of the Garden of Eden instead of somewhere beyond reach? Why did he put it there in the first place? Why the snake? Why the seduction? What were God’s intentions? Why does He ask for sacrifices? Why sacrificing His son? Why does He not stop evil Himself? Why does He work in such mysterious ways? How do we know which parts of the Bible should be interpreted only metaphorically? And so on and on.

The two principles of translation invariance and context independence imply that we can use instruments: we can use a ruler, or we can use something to cover or erase the arrowheads. With these instruments, we can clearly demonstrate that both horizontal lines are equal. Our two universalized principles cohere with each other. But they are in contradiction with this one intuitive judgment about the different lengths of the horizontal lines.

We now have two options. 1) We can abandon two of our strongest and coherent intuitions (translation invariance and context independence) and try to make a consistent geometrical system without those two principles, in order to save our intuition that the horizontal line segments are of different lengths. As mathematicians often invent some very exotic geometrical systems, this strategy is not necessarily impossible. But everyone would agree that such a procedure to invent a new geometrical system would be very difficult, and the resulting geometrical system will appear to be very artificial. 2) So a second option is to acknowledge that our intuition about the lengths of the horizontal lines is wrong. The intuition does not disappear however; it has some cognitive impenetrability, so it is not a mistake or deception. It is an illusion.

Most people automatically prefer the latter option, because the combination of the intuitions of translation invariance and context independence is very strong, and we do not want to dismiss them so easily. Translation invariance and context independence generate a coherent system in narrow reflective equilibrium (Daniels, 1979).

To this coherent system we can add some background theories to arrive at a wide reflective equilibrium. Two background theories are added: one from psychology, one from anthropology. Adding knowledge about the underlying psychological mechanism generates more evidence that we are dealing with an illusion. We know that the Müller-Lyer optical illusion is created by our brains in order to adapt a 2D retinal image to 3D-vision. Two objects of equal physical length can have different images on our retina, if one object is further away from our eyes than the other. Our brains correct for this difference in appearance. So the mechanism is perspective-adaptation. As we live in a 3D-world, our brains are trained to make 3D-adaptations. Sometimes they're stuck when looking at a 2D-image such as the Müller-Lyer image.

What actually happens can be described as a psychological heuristic (Kahneman & Shane, 2002). Heuristics are intuitive, efficient rules of thumb that are applied when facing complex problems. They work by a process of attribute substitution: a target attribute that is computationally complex for our brains is (unconsciously) substituted by a heuristic attribute that is easier to calculate. Kahneman (2003) argued that some optical illusions are generated by heuristics and attribute substitutions. The following image of a staircase clarifies this a bit more.

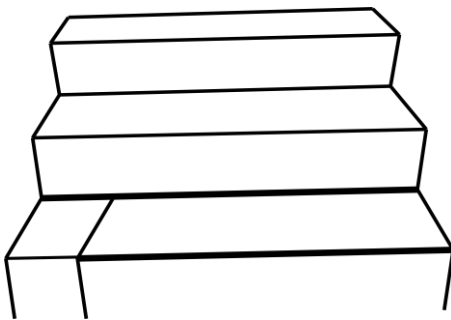


Figure 2: the Muller-Lyer optical illusion in a staircase

In 3D, we know that each step has the same size, so each horizontal line has the same length in a 3D-world. But in 2D, the lines have clearly different lengths. Now we see this picture in 2D, and we have to determine the lengths of the two thick horizontal lines of the bottom stair. These lengths are the target attribute. But what do our brains do? We often take the stairs, so our brains are used to computing sizes of objects in 3D. Sizes in 3D are therefore easily accessible, and our brain unconsciously uses them as heuristic attributes to determine sizes in 2D-pictures. That is how length judgments in 2D-

pictures get distorted: the lengths of the two thick lines are, in fact, equal. But in a 3D staircase, the lower thick line would be much shorter, so that's why it appears shorter.⁵

A second background theory refers to anthropology. Interestingly, the Muller-Lyer illusion is not inborn. Anthropological studies have shown that the illusion depends on culture (Segall, 1963; Ahluwalia, 1978). In particular, some indigenous people (who do not live in an environment with straight edges of houses, tables and staircases) are less susceptible to this optical illusion. So if they don't see it and we do, who is right? This is another part of the evidence that we are indeed dealing with an optical illusion.

Two principles that cohere with each other, added with background theories about the psychological mechanism and the cultural relativity, together form quite some evidence to justify the claim that it is an optical illusion. The only counterevidence is that the illusion does not simply disappear after reflecting about it.

Let's consider a second example of an optical illusion: the grating induction illusion (Foley & McCourt, 1985).

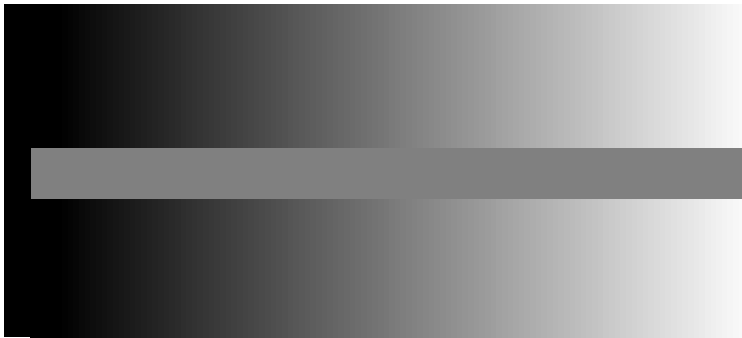


Figure 3: the grating induction illusion

The horizontal grey bar is actually uniformly grey, but the left side appears brighter. We can check translation invariance by taking a piece of paper with the same grayness as the left side of the bar, and then shift the piece of paper to the right side. Our intuition says that the properties of paper (its grayness) do not change when the paper is shifted. We can also check context independence by covering the black-white areas around the grey bar. Our intuitions say that grayness of an object does not depend on things around it. As a third argument, we know that our optical system is also equipped with a mechanism of lateral inhibition, in order to increase the contrast and sharpness of visual stimuli. This mechanism generates the grating induction illusion. So we understand where the illusion comes from.

⁵ For a more detailed explanation of this illusion, based on a mechanism of time-delay, see Changizi et al. (2008).

In summary: two coherent arguments based on strong intuitions of invariance and context independence (non-arbitrariness and non-artificiality), and a supporting argument that says something about the underlying mechanism in our optical system, gives us enough evidence to counter a weak perceptual intuition. Can we apply the same method to moral intuitions?

3.2 Moral illusions

So let's turn to ethics now. Moral illusions are obstinate but incorrect intuitive judgments, comparable to the famous optical illusions. Sir David Ross (1930) compared our moral convictions or intuitions with sense-perceptions: the former are the basic data of ethics, just like the latter are the basic data of natural science. But he remarked: "Just as some of the latter have to be rejected as illusory, so have some of the former; but as the latter are rejected only when they are in conflict with other more accurate sense-perceptions, the former are rejected only when they are in conflict with other convictions which stand better the test of reflection." (Ross, 1930, p41) By making this analogy with sensory illusions (e.g. optical illusions), Ross might be one of the first ethicists to point at the existence of moral illusions.

Seventy years later, as scientists and philosophers became more and more interested in the neurobiology of morality, the notion of moral illusions and its analogy with optical illusions reappeared. One philosopher and neuroscientist made the point explicitly clear: Sam Harris stated that, for example, the difference in moral disapproval between torturing a suspected terrorist (to find the location of a bomb that is about to kill hundreds of people) and collateral damage in war is a moral illusion. "Paradoxically, this equivalence [between using torture and causing collateral damage] has not made the practice of torture seem more acceptable to me [...]. I believe that here we come upon an ethical illusion of sorts – analogous to the perceptual illusions that are of such abiding interest to scientists who study the visual pathways in the brain. The full moon appearing on the horizon is no bigger than the full moon when it appears overhead, but it looks bigger, for reasons that are still obscure to neuroscientists. A ruler held up to the sky reveals something that we are otherwise incapable of seeing, even when we understand that our eyes are deceiving us." (Harris, 2004, p198) Harris continues with pointing at a possible psychological bias behind our moral intuitions: "In fact, there IS already some scientific evidence that our ethical intuitions are driven by considerations of proximity and emotional salience of the sort I addressed above. Clearly, these intuitions are fallible. In the present case, many innocent lives could well be lost as a result of our inability to feel a moral equivalence where a moral equivalence seems to

exist. It may be time to take out our rulers and hold them up to the sky.” (Harris, 2004, p198)

Two questions need to be answered. First: do moral illusions exist? Second, and more important, how do we know? How can we agree whether some moral intuition is or is not a moral illusion?

To answer the first question, let us again look at the six analogies of ethics. Do these other fields of cognitive activity contain illusions as well?

1) In modern physics, we encounter contra-intuitive judgments in e.g. quantum mechanics and relativity theory. We could say that some intuitions about simultaneity, measurements, particle identity or space-time are illusions.

2) In mathematics and statistics, too, we encounter erroneous intuitive judgments such as in the famous Monty-Hall problem or the mysterious Banach-Tarski property. Studies on heuristics and cognitive biases (Kahneman et al., 1982) show that our intuitive judgments under uncertainty (e.g. in statistics) are not always reliable.

3) In the field of geometry we have the optical illusions.

4) In grammar we have an interesting situation. We already mentioned the apparent analogy between our grammatical faculty and our moral faculty. If moral illusions would exist, then grammatical illusions might exist also. And this is indeed the case. (Phillips et al., 2010) A simple example of a grammatical illusion is the sentence: “One out of three children are overweight.” According to a lot of people, this sentence appears to be grammatically correct at first sight. Yet, it is a violation of a most simple rule of subject-verb agreement. The fact that people repeatedly make such errors can indicate that it is an illusion instead of merely a mistake. Such errors are too persistent to be merely mistakes.

5) Our example of crossword puzzles might perhaps be too rudimentary to have illusions. (Mistakes, however, are often made in solving crossword puzzles. But as we have seen, illusions are more persistent than mistakes.)

6) In our judgments about taste preferences, there can also be deceptions (if you like apples and I give you a piece of apple, paint it with a brown, odorless, tasteless color and cut it in the shape of a sausage, then you might judge it to be bad) as well as illusions. For example, psychological biases can influence our taste preference. Taste evaluation is influenced by what we think we eat and whether that food symbolizes values that we support. For example meat eaters and vegetarians are susceptible to this kind of taste illusions. Researchers (Allen et al., 2008) have given meat eaters two sausages. The participants thought that the first sausage was meat (whereas in reality it was a vegetarian sausage) and that the second sausage was vegetarian (whereas in reality it contained meat). The meat eaters who valued dominance, hierarchy and social status tended to prefer the first sausage, because their taste preference is influenced by their value scheme. Taste evaluation is not simply a matter of chemistry.

Also in other fields illusions do exist. There are auditory illusions (e.g. a pitch seems to increase indefinitely; Deutsch, 1992), sense illusions (e.g. the contrast effect: place your left hand for some minutes in cold water and your right and in warm water, then touch with your both hands the same object) and many more. Illusions appear frequently in different areas, and morality is not likely to be an exception.

Just as optical illusions can learn us a lot about how our visual perception system works, and grammatical illusions can inform us a lot about how our language faculty works, so could we learn a lot about morality by focusing at moral illusions.

After having affirmatively answered the question whether moral illusions might exist, let's now move to the second, more interesting question: how do we know whether a moral intuition is an illusion?

One might think that with optical illusions matters are easy, because there is an objective reality to refer to. With ethics, we do not have such an objective reality. However, in the case of the optical illusion, we do not need to get a direct access to an objective reality. In fact, we can use some non-argued basic starting points or intuitions, such as translation invariance and context independence. The underlying intuitions behind the principles of translation invariance and context independence lack further foundational justification. These are principles which we have to agree to accept.

So how to tackle moral illusions? We in fact already have the answer. We start from moral intuitions, because there is nothing else to start from. These intuitive moral judgments in particular situations have to be expressed in particular ethical rules. In the next step, these particular ethical rules have to be universalized to all other similar situations. After formulating universal ethical principles, we have to check whether the resulting system has internal consistency. By testing more and more situations, i.e. by looking whether our moral intuitions in all situations agree with the universalized ethical principles, we get a coherent system. If we arrive at a contradiction, i.e. if a moral intuition is incompatible with our universal ethical principles, we could refine some principle or introduce a new principle that settles the conflict. If this strategy really doesn't work, then the only option left is that our moral intuition is wrong. If this intuition does not disappear, we have found a moral illusion.

In the Müller-Lyer illusion we had reliable instruments to demonstrate that it is an illusion: we need a measuring stick or something to cover or erase the small arrowheads. In ethics, our reliable instruments are valid arguments based on universalized principles coming from strong intuitions in reflective equilibrium. So, valid arguments are the reliable instruments to demonstrate that a moral intuition is an illusion. (Yet, in the coherence picture, nothing has absolute reliability. Even the strongest intuitions can be mistaken, even measure sticks might be untrustworthy, even basic experimental data in science might be erroneous. The strength of arguments or principles lies in the overall coherence.)

There is one feature that almost all optical illusions have: the influence of context. In the Müller-Lyer figure, the small arrowheads are the context. Our geometric system requires some context independence, because the context is irrelevant and arbitrary. Yet, the context might influence our perceptions and judgments. This might also be the case with moral illusions. We can expect that moral illusions are to be recognized by their context dependence, arbitrariness, artificiality, complicity or fuzziness.

The idea of moral illusions sheds a new light on the problem of moral disagreement. Giving us tools to demonstrate that a moral intuition is in fact an illusion will help us to accept a coherent theory of e.g. animal equality. In geometry we had tools to demonstrate that an intuition is an illusion. We now have similar tools in ethics: strong moral intuitions that can be translated into coherent universalized ethical principles.

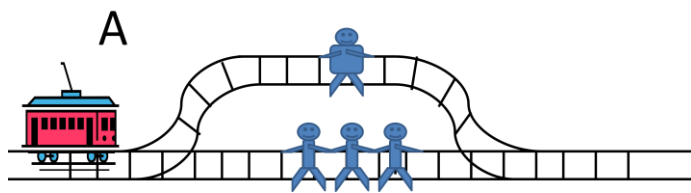
So the analogy between optical and moral illusions can help us to better detect and understand moral illusions. The strategy for detecting optical illusions – using translation invariance, context independence and an underlying optical mechanism – can be applied to ethics as well, in order to detect moral illusions.

There might be quite a lot of examples of moral illusions: perhaps futility thinking and projective grouping (Unger, 1996, p.100), moral luck (Nelkin, 2013) or the intransitivity problem (Temkin, 1987) are examples of moral illusions. In recent literature, as a spin-off of the work of Kahneman & Tversky (1982), the study of moral heuristics gained some influence (Sunstein, 2005; Sinnott-Armstrong et al., 2010). As Sunstein argued, in certain situations, moral heuristics might create erroneous intuitive judgments that we could also consider as a specific kind of moral illusions. In a later chapter, I will look at a more debated issue, the prejudicial difference in moral status between humans and non-human sentient beings (even when the humans have equal levels of mental capacities as the non-human animals). This is to most people an intuitive judgment that is used to justify all kinds of (ab)uses of animals, from factory farms to pet shops. Can this speciesist intuition be a persistent moral illusion? Is it an erroneous moral heuristic? Before we move to the animal issue, it might be a good idea to first apply our “moral illusion detection technique” to the psychological mechanism of intervention myopia (Waldmann & Dieterich, 2007) in the trolley problem.

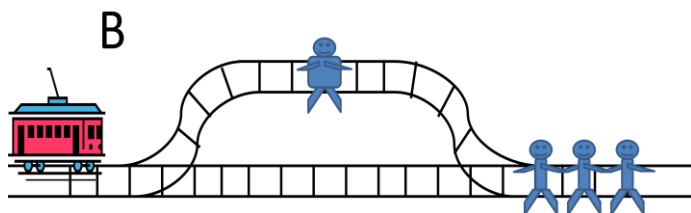
3.3 An example of moral illusions in the trolley dilemma

Let's discuss a moral illusion in some detail. Consider the following trolley dilemma (based on Thomson, 1985; for a review on the trolley problem: see appendix). A runaway trolley is moving on the main track, endangering three people. You can save those people by turning a switch so that the trolley takes a side track. However, on this side

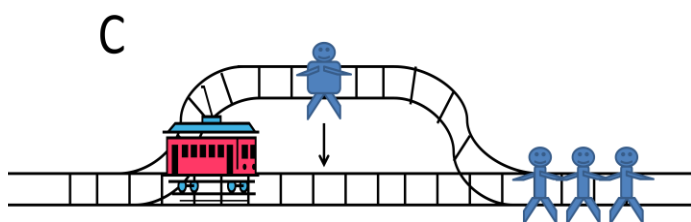
track there is one person who will be killed. In summary: doing nothing results in the death of three persons, acting (turning the switch) results in the death of one person.



In situation B, we change the set-up a little bit. This time, the three people on the main track are behind a second fork. Again, doing nothing results in the death of those three people. But turning the switch redirects the trolley to the side track, on which there is a heavy man, who weighs enough to block the trolley. This is the loop trolley dilemma, because the side track loops back towards the main track.



Situation C is again slightly different. This time, you hesitated too long and the trolley already passed the first fork, heading along the main track towards the three people. The plan to turn the switch will no longer work. But the side track is on a bridge above the main track. So you could still stop the trolley by pushing the heavy man from the bridge in front of the trolley on the main track.



As you can see, the differences in the three dilemmas consist in the positions of the trolley and the three people on the main track: they can be either before or behind the first and second forks.

According to psychological studies (Hauser et al., 2008), roughly 90% of the people respond that turning the switch is allowed in the first dilemma (situation A). Only 50% say that we are allowed to turn the switch in the second dilemma, whereas 10% state that we are allowed to push the heavy man from the bridge in situation C. By looking at

brain activity, Greene et al. (2001) discovered that persons reacted differently (more emotionally) in the third dilemma (situation C), compared to the first. Other brain regions became active.

We can give two differences between the above three trolley dilemmas. The first distinction separates situation A from situations B and C. In situations B and C, looking at the trajectory that the trolley will take, we place the heavy man in between the trolley and the three people. That is necessary, because the heavy man has to block the trolley. If the heavy man was not present, the trolley could move on and kill the three people. So if your plan is to save those three people, the presence of the heavy man is required in order for your plan to work. This first principle can be expressed as a deontological right not to be used as merely means. The heavy man is used as merely means to save the others. It is a deontological right, as it can be in conflict with a consequentialist right to live. According to consequentialism, it is always morally required to act in all three dilemmas, because the consequentialist right to live is violated only once by acting, whereas it is violated three times by allowing the three people on the main track to die. If the deontological right is at least three times stronger than the consequentialist right, then action is not permissible in situations B and C. In those situations, one deontological right is violated.

A second distinction, that separates situation C from situations A and B, is that in situation C, the action is up-close-and-personal. You have to touch and push the heavy man yourself. But suppose you simply have to push a button from a distance, to overturn the side track and drop the heavy man on the main track. I expect that even then a majority of people (more than the 50% of people who responded favorably in situation B) would respond that action is not permissible. This expectation is compatible with a second psychological study on the trolley dilemmas (Waldmann & Dieterich, 2007).

In that study, all people on the main and side tracks were actually sitting in busses. Situation C then consists in pushing the bus from the side track onto the main track: you are sitting in a heavy truck that can push the bus, so no personal contact is required. As the person on the side track is sitting in a bus, not he, but the bus is used as merely a means. If the person was not in the bus on the side track, your plan to block the trolley by using the bus would still work. So the death of the person in the bus could be considered a side effect.

As the deontological right is not violated in the bus-trolley dilemmas studied in Waldmann & Dieterich (2007), we expect that people are more permissive towards action, and that there is no distinction between situations A and B (because the only distinguishing factor was the violation of the deontological right). This is indeed the case. Respondents could rate the admissibility of action from 1 (definitely not allowed) to 6 (definitely allowed). The bus-trolley dilemmas A and B received an average rate of about 4,8, without a statistically relevant difference between the two dilemmas. So

action in both situations is allowed to a high degree. But situation C received an average rating of 3,7, which is relevantly lower than the other two situations.

As in situation C there is no longer a personal contact with the victim, there is something else at hand. The difference between situation C and situations A and B is that in C the victim (the heavy man on the side track) is sent to the threat (the trolley), whereas in A and B the threat is directed towards the victim. In other words, looking at the causal chain in situation C, the 'locus of intervention' is at the person on the side track. Sending the victim to the trolley means that the path of the victim is influenced directly. On the other hand, turning the switch in situation B means that the locus of intervention is at the threat. Sending the trolley to the side track means that the path of the threat is influenced directly (and the path of the victim is influenced indirectly because the threat will eventually hit him). So the difference between situations B and C is the causal path and the locus of intervention. Directly intervening in the path of the victim in situation C is a generalization of the up-close-and-personal element of pushing a heavy man from the bridge. There is no close contact with the victim, but still there is some directness.

According to the study of Waldmann & Dieterich (2007), direct intervention in the path of the victim is considered less permissible than intervention in the path of the threat, even in the absence of up-close-and-personal contact. And according to the study by Hauser et al. (2008), 50% of respondents said that action in situation B is not allowed. From these facts, we can expect that a majority of people (significantly more than 50%) would respond that action in situation C is not allowed even in the absence of up-close-and-personal contact. With close contact, 90% of respondents said that action is not allowed in situation C.

Note that the results in both studies are not contradictory either, because in Waldmann & Dieterich (2007) the situations contained no violations of deontological rights, and in Hauser et al. (2008) there was a very close contact with the victim in situation C.

In summary, we have two distinctions between the three trolley dilemmas.

1) Action in situation A implies no violation of the deontological right not to be used as merely a means. In B and C this right is violated. This distinguished A from B and C.

2) Action in situation C implies that the victim is sent to the threat. In A and B, the threat is sent to the victim. This distinguishes C from A and B.

The first principle says that action is less permissible if a deontological right not to be used as merely a means is violated. The second principle says that action is less permissible if the victim is sent to the threat. We have seen that these principles correspond with two psychological studies on the trolley dilemma. In situation A, action means that the deontological right is not violated and the victim is not sent to the threat, so it is strongly permissible (as 90% of people say). In situation B, action still means that the victim is not sent to the threat, but the deontological right gets violated.

So this is much less permissible (only 50% of respondents say it is permissible). In situation C, the victim is used as merely a means and is sent to the threat. So in this case, the situation is strongly impermissible (as a majority says).

The question we now have to ask is: are the moral intuitions behind these two principles illusions? I will demonstrate that the second principle is based on a moral illusion, but the first one is not. To show that the second principle is a moral illusion (as was hinted at by Peter Unger, 1996), I present (just as with the optical illusions) two arguments based on strong moral intuitions, and one auxiliary argument based on knowledge of the underlying psychological mechanism.

The first argument is based on a kind of translation invariance. So we have to see what remains constant as we shift from the first to the second and the third situation. In the Müller-Lyer optical illusion, we have seen that what remains constant when shifted from the upper to the lower part of the figure, is an intrinsic property of a line segment: the length. The length of a ruler does not change when shifted. In the trolley dilemmas, what remains constant are intrinsic and morally relevant properties of the people involved. The moral status of an individual is such an intrinsic (context independent) and morally relevant property.⁶ When moving from dilemma B to C, the moral status of all individuals remains the same, it is independent from the situation. This seems self-evident, but is in fact a strong intuition that we strongly accept. This moral status can consist of different things, such as the consequentialist right to live and the deontological right not to be used as merely a means. These are rights that individuals have, no matter what the situation may be. Most importantly, we have a strong intuition that the moral status of an individual does not change when the locus of intervention in the causal chain is changed. The moral status of a victim is not higher when the locus of intervention is at the victim instead of at the threat.

The second argument is based on context independence. In the Müller-Lyer illusion, we can cover or erase the irrelevant properties, such as the arrowheads, to check the equality of lengths. So looking at trolley situations B and C, what is the irrelevant context that we have to erase? The only relevant aspect is the collision between the trolley and the victim. Where this collision happens is not relevant. So let's simply erase the tracks and all other things in the environment. We are left with a completely empty

⁶ With this I mean that an individual has objective properties (such as mental capacities for well-being and consciousness) and that we can attribute a moral status to these objective properties. The moral status itself is not an objective property (not a moral fact, as cognitivists would say), but an attributed property (comparable to secondary qualities such as color). In this sense, color illusions (such as the grating induction illusion) might be a better analogy of moral illusions than the Müller-Lyer illusion, because length is a primary quality whereas color (greyness in the grating induction illusion) is a secondary quality. For the analogy in meta-ethics between moral values and secondary qualities, see e.g. McDowell (1984).

space, apart from the two relevant entities: the trolley and the victim (the heavy man). In situation B, the trolley is moved upwards, towards the heavy man. In C, the heavy man is moved downwards, towards the trolley.



As there is no absolute point of reference around (we have erased all irrelevant things), relativity says that both situations B and C are now in fact equal. According to Unger (1996, p101), this difference between sending a victim to a threat and a threat to a victim is based on what he called ‘protophysics’, in violation of relativity theory. This demonstrates that the difference between situation B and C is simply something contextual, as the context determines the spatial frame of reference. But of course, this argument is just as well based on (strong) intuitions about what is the context and what is morally (ir)relevant. The position of the tracks is not morally relevant. The only relevant thing is the relative position of the threat and the victim, because this determines the collision.

The third, auxiliary argument is based on a psychological mechanism. This mechanism is in fact clearly explained by Waldmann and Dieterich (2007). Their concept of ‘intervention myopia’ already indicates that we are dealing with something that is not functioning properly. When the locus of intervention is at the threat, as in situations A and B, our attentional spotlight is at the threat, and all people (on the main and the side track), are background. So from the perspective of the threat, the persons are all equal, and it is easier to make consequentialist calculations. But if the locus of intervention is at the victim (the heavy man), the focus of attention is at the victim. The other three people on the main track are now part of the background. And the myopia indicates that it is difficult to take those people in the background fully into account. Due to the myopia, we don’t see their moral status so clearly. This distorts consequentialist reasoning. In situation C, people tend to focus on the fate of the victim, neglecting the death of the three people on the main track. This focus on the one victim (the locus of intervention), results in a neglect of other people located in the background. As these three people in the background appear to be absent (far away in the causal chain), it appears that the heavy man dies in vain when pushed from the bridge. Citing Waldmann & Dieterich: “In sum, the general hypothesis is that people tend to focus on the causal paths of agents [threats] or patients [victims] targeted by an intervention, and neglect other causal processes occurring outside this focus, in the background.[...] We are not saying that in cases of intervention myopia, people are completely blind to the victims in the background (i.e., the death of the three people);

rather, we are saying that because of an attentional focus on the effects of interventions, people who are evaluating the morality of options may give victims in the background less weight than victims in the attentional spotlight.” (Waldmann & Dieterich, 2007, p249)

With the above arguments, we can conclude that the difference between sending a threat to a victim and sending a victim to a threat is a moral illusion. The moral relevance of this difference was already criticized by Fischer (1992) and Fischer & Ravizza (1994), but now we have a clearer view and we can call it a moral illusion. Whether the deontological right not to be used as merely means is also a moral illusion, remains to be seen.

3.4 Is the deontological right a moral illusion?

As we have seen in the discussion of the trolley dilemmas, there is a possible explanation that separates situation A from situations B and C. In situation A, a deontological right of the victim is not violated. In situations B and C, the victim is used as merely a means, as a trolley blocker or a human shield. The presence of this victim is required in order to save the other people on the main track.

As far as I know, the coherentist approach does not yet imply that this deontological right is a moral illusion. On the one hand, the above mentioned intervention myopia seems to suggest that the deontological right is an illusion. But on the other hand, there seem to be some arguments that suggest that the deontological right is not an illusion.

First, the deontological right respects a translation invariance: all persons keep the same deontological right in any of the trolley dilemmas, because the deontological right is related to the moral status of the individual, and this moral status is invariant (it remains the same when shifting between different dilemmas).

Second, there seems to be a context independence. It can be said that all people can claim this right, independent from the situation. So it has some intrinsic (context independent) character.⁷

Third, and perhaps most importantly, there are hundreds of other moral dilemmas and situations where intuitive judgments of most people are coherent with the

⁷ However, this right refers to the use as merely a means to someone else's ends, so it refers to the presence of someone else in the environment. It is unclear whether this introduces a context dependence and whether this kind of context dependence creates an illusion.

deontological right principle. We are not allowed to sacrifice an innocent person against his will to use his organs to save five patients in the hospital who need new organs in order to live. Neither are the following actions allowed: involuntary experiments, terror bombing (killing innocent civilians in order to demoralize the enemy), torturing a suspected terrorist (to gain information on the location of a bomb that is going to blow up a school), killing under blackmail (a terrorist says that if you kill an innocent person, he will not kill his five hostages), trafficking (buying and selling humans), raping women (and selling the video to thousands of male consumers), gladiator fights (entertaining thousands of spectators), human exhibitions (ethnographic zoos), cannibalism or slavery. In all these cases, the presence of the victim is required in order to benefit others, so the victim is used as merely a means. The above actions are not allowed, even when consequentialist considerations might support those sacrifices and deontological rights violations.

So we have more than ten dilemmas and situations where the deontological right might be violated: trolley-bridge, organ transplants, experiments, terror bombing, torture, blackmail killing, trafficking, rape, cannibalism, gladiator fights, and slavery. In all these different situations, most people have the very coherent intuition that the deontological right should be protected against consequentialist considerations. If the deontological right not to be used as merely a means trumps the right to live, then we are not allowed to use people as e.g. trolley blockers or information sources, even if this means that other people will die.

As the deontological right can be related to the moral status of an individual independent from the situation, and as there are a lot of very different situations where most people's moral intuitions are coherent and compatible with the principle of the deontological right, I am tempted to believe that this right is not (yet) a moral illusion.

The only exception to this deontological right intuition known to me is the loop trolley dilemma. The loop dilemma (situation B) is often mentioned as a counter-example to the moral relevance of the deontological right (e.g. Singer, 2005). In the loop dilemma, the deontological right is violated, yet, the action is deemed permissible by a lot of people (Hauser et al. 2008). Consequentialists can refer to this intuition in this particular dilemma to argue that the deontological right is a moral illusion. But from the above discussion, we learn something interesting: the permissibility in the loop dilemma might be a moral illusion, and if this is the case, then the deontological right might still be valid. So the existence of moral illusions might 'save' the deontological right.

Now, if the deontological right is not an illusion, what happens with the above psychological explanation of intervention myopia? In short, we might hypothetically say that intervention myopia does not generate the illusion in the bridge dilemma, but 'instrumentalisation myopia' generates an illusion in the loop dilemma. The structure of the loop dilemma is such that we don't see the instrumentalisation (the use as merely a

means) of the person on the side track, because that person seems as far away as the non-instrumentalised person on the side track in dilemma A. In other words: the person on the side track is too far away for us to see his instrumentalisation.

In terms of heuristics, the attribute substitution might (very hypothetically) work as follows. The target attribute is in this case the deontological right. As it is not always easy to quickly detect violations of this right, our brains might use a heuristic attribute instead. Note that, if the deontological right gets violated, it means that the presence of the victim is required in order to save other people. The required presence is the target attribute, and it requires a sometimes computationally difficult counter-factual thought experiment to determine whether someone's presence is required. But when presence of the victim is required, this likely means that the focus of the action (the locus of intervention) will be on the victim. In other words, the locus of intervention might be a good heuristic attribute for detecting deontological rights violations. Therefore, our brains look for this heuristic attribute, which is reliable in most cases, but misfires in the loop dilemma. In the loop dilemma, the locus of intervention was not on the victim, and hence our brains think erroneously that the deontological right is not violated.

The above is still very hypothetical, because one might also say that the true target attribute is the maximization of well-being or lives saved, as consequentialists would have it. So the issue remains open: is the deontological right (the intuition that inhibits action in the bridge dilemma) a moral illusion in a consequentialist ethical system, or is the loop dilemma intuition (that action is permissible) a moral illusion in a deontological system?

Whether the deontological right not to be used as merely a means is also a moral illusion, remains to be seen. As the deontological right, due to its intrinsic and context independent character, can be related to the moral status of an individual independent from the situation, and as intuitions in a lot of different situations are coherent with the principle of the deontological right, I am tempted to believe that it is not a moral illusion. This deontological right will be discussed in more detail in the later section on the basic right (section 6.2).

3.5 Heuristics in thought experiments

Sunstein (2005) criticized the above method of coherentism (reflective equilibrium) that uses philosophical thought experiments (moral dilemmas like the trolley problem). His claim is that our intuitions in those 'exotic' (far-from-ordinary-reality) thought experiments are not reliable anyway, because exotic situations are often situations where heuristics misfire. Heuristics are common sense rules of thumb that work well in

most casual situations that we encounter in our daily lives, because in these situations we are trained to make quick and accurate judgments. But in constructing exotic thought experiments with trolleys, we create situations where the heuristic does not yield reliable results, i.e. where the heuristic attribute strongly deviates from the target attribute. At least that is what Sunstein claims.

Yet, I would not throw away the method of thought experiments. I claim that those thought experiments do have some value: they trigger our moral intuitions and interrogate our morality, just as experiments in physics allow us to interrogate nature. I'd like to defend the coherentist-universalist approach (deriving intuitions from thought experiments and translating them in coherent universalized ethical principles), by making the analogy with physics. For example, in our daily lives, we often see that heavier objects fall at higher speeds. Yet, the intuition that heavier objects have higher accelerations deviates from the real law of gravity. So in order to derive the laws of gravity, we have to set up exotic situations, e.g. by dropping different objects in a vacuum (e.g. on the moon). Those exotic experiments are controllable, they eliminate specific contextual variables (like air draft) so they are better suited to interrogate nature to find its most fundamental laws. And the same goes for ethics, by interrogating our moral brains, using exotic thought experiments, we can derive the fundamental laws, the basic forces of our ethics.

Yet, even in the exotic experiments, we cannot trust everything. In physics, to derive the acceleration law of gravity, we have to use clocks. But as Einstein demonstrated in his theory of general relativity: clocks measure different times depending on their positions in a gravitational field. So clocks can move faster or slower. Hence, even in exotic experiments, we cannot completely trust our instruments. In our daily lives, clocks are reliable to measure time, so we use clocks as heuristics. The value on the clock is a heuristic attribute; the real time is the target attribute. Now, in some exotic situations, with strong gravitational fields, the heuristic does not measure real time anymore. Consider clocks in GPS satellites. A clock in a GPS satellite experiences a weaker gravitational field than clocks on earth, so a satellite clock runs at a slightly higher speed than a similar clock on earth. We can use the clock on earth as a heuristic to measure time evolution in the universe. But this heuristic misfires (a tiny bit) when we want to measure time evolution in the satellite. We know, thanks to Einstein, that clocks on earth might not be reliable heuristics in those exotic situations. The only way to discover this heuristic misfiring is by setting up other exotic experiments, and deriving a coherent framework about how time evolves at different places in a gravitational field. That coherent framework was derived by Einstein based on (thought) experiments.

So, indeed, we cannot always rely on moral heuristics in exotic situations. Indeed, exotic situations might sometimes be exactly those situations where heuristics misfire. But using more and more thought experiments, we can see when our heuristics might

misfire. They misfire when they are not consistent with a coherent ethical theory of universalized ethical principles.

There is another, slightly related issue. We construct thought experiments by erasing a lot of variables. Like the trolley dilemma, those experiments are characterized by their low number of variables. E.g. we don't consider interpersonal relationships, probabilities or people's virtues in the trolley dilemma. This helps us to look for those elements that are crucial in our study. However, one critique can be that they neglect the specificities of contexts and particular, real situations. Some complexity and situation-dependence is missing.

Here we can use another analogy with physics. One branch of physics is looking for the most fundamental laws: high energy physics (elementary particle physics and cosmology). Other branches study other physical contexts: for example fluid mechanics or statistical thermodynamics. Although a fluid is composed of elementary particles subject to the fundamental forces, fluid mechanics studies other, emergent laws and principles.

In this dissertation, I intend to derive fundamental laws of ethics, the basic 'moral forces', so it is comparable with elementary particle physics. I use thought experiments in ethics, just as particle physicists perform exotic experiments using e.g. particle accelerators. Contextualist or situationist ethics, on the other hand, is most suitable for complex real life issues. These ethics might be comparable to e.g. fluid mechanics or thermodynamics. In these contextualist ethics, new rules or principles might emerge, that strongly depend on the specific context (e.g. complex relationships or cultural influences). The study of elementary particles is not incompatible with the study of fluid mechanics. And for the same reason, a principle-based, universalist ethics is not incompatible with more contextualist ethics, as long as the laws of the latter are emergent from the laws of the former. Different contexts (cultures, relationships,...) might require different emergent rules, just as fluids, gases and solid states have different emergent properties.

Summary of part one

In this chapter I have argued that moral intuitions are the starting points. These intuitions are spontaneous judgments lacking further justification. The method of ethics consists of articulating these moral intuitions into particular ethical rules and then universalizing these particular rules with respect to the act, the situation, the moral agents and the moral patients. Different universalized principles might cohere with each other, giving us more and more confidence in their validity. At the end we arrive at a coherent reflective equilibrium, an ethical system that is consistent, clear, parsimonious and in agreement with our strongest moral intuitions. If there remains an intuition which is in contradiction with the strongly coherent system (which cannot be incorporated in the system) and which does not seem to disappear, we have to admit that this intuition is a moral illusion.

Let us visualize this process with the following figures. Figure 4 represents the starting point. The grey areas represent moral intuitions that we derived from looking at different moral dilemmas. The white areas are unexplored. The question is how to fill them in.



Figure 4

Let's start with the bottom part of the figure. There are two grey beams. The simplest way is to connect them with each other. In other words, we extend the grey beams into the white area. This is a process of universalization: all areas in the figure with the same level of grayness reflect all situations with the same moral judgment. Compare it with a

crossword puzzle, whereby we have already found the following letters: UNIVER
ZATION. We can now easily complete this word. This gives us the next figure.



Figure 5

Note that I also erased the two black spots, because they were a bit ‘ugly’ looking and didn’t fit the picture. These stains were also easy to erase so they are the analogs of moral mistakes or deceptions.

In the next step, we could extend (universalize) the grey area to the middle of the figure. This gives us the next picture.



Figure 6

The question now is how far to the left can we extend the grey beam. How far do we have to universalize the principle? Should the left part be black or grey? We could invent a moral dilemma that represents the left area in the figure, and test our moral intuition in that situation. We find out that our moral intuition in that dilemma indicates that the area should be black. The next picture shows that this is coherent. We had the intuition that it should be black, so the black vertical beam on the left can be universalized, and we see that the grey beams at the bottom and the middle of the figure are surrounded with black in a similar way. This confirms our solution.



Figure 7

And finally, we could fill in the remaining white areas by simply ‘universalizing’ (extending) the grayish areas in a consistent way. The next figure shows the end result.

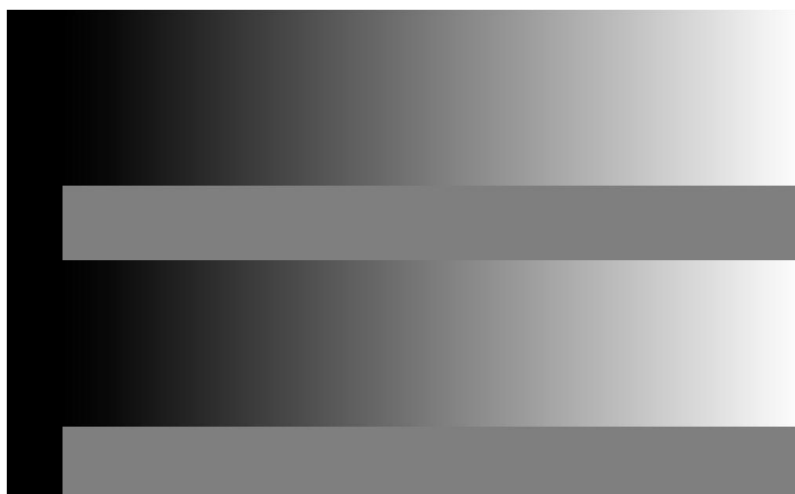


Figure 8

But now something strange happens: a moral illusion appears. It seems as if the intuitive judgments represented by the left side of the beam are different from the moral judgments of the right side. Nevertheless, our system is coherent.

We have seen a method to check whether an intuition is an illusion: the process of moving from our intuitions towards a system of universalized ethical principles in a coherent reflective equilibrium. This is also the process that people follow in empirical sciences, mathematics, grammar or taste evaluations. We can discover deceptions and illusions, even in the area of taste preferences. So rational arguments are possible, at least to some degree.

Four recent developments indicate that the time is ripe to tackle moral illusions. First, a few philosophers started comparing erroneous moral judgments with optical illusions (Harris, 2004). Second, there is the study of ‘moral heuristics’ (Sunstein, 2005;

Sinnott-Armstrong et al., 2010) and the discovery that attribute substitution is also the mechanism behind many optical illusions (Kahneman, 2003). Third, there is the study of ‘grammatical illusions’ (Phillips et al., 2010) combined with the linguistic analogy between our moral faculty and our language faculty (Hauser et al. 2008). Finally, there is the (neuro)psychological research of the trolley problem (Greene, 2002; 2008), with the hypothesis of intervention myopia in moral intuitions (Waldmann et al. 2007). The analogy between optical, grammatical and moral illusions can help us to better detect and understand moral illusions.

In a later chapter, I want to focus at a specific moral judgment that meat eaters have: the prejudicial difference in moral status between humans and animals (non-human sentient beings). This is to most people an intuitive judgment that lies behind all our uses of animals, from factory farms to pet shops. Can this intuition be a persistent moral illusion? I will demonstrate that this speciesist judgment is in fact a moral illusion, just like I demonstrated that the Müller-Lyer figure is an optical illusion and the trolley dilemmas contained a moral illusion.

In particular, the set of universalized ethical principles that I will discuss contain two principles based on strong moral intuitions of invariance and context independence. And as an auxiliary argument, I will explain the psychological mechanism behind the moral illusion of speciesism.

As the purpose of this dissertation is to investigate the consistency of an ethic of animal equality, the most important thing is that both animal rights ethicists and meat eaters should agree on the approach: they should both agree on the key ingredients, the moral intuitions as input data, universalization as the method, and coherence or consistency as the goal. If we agree upon an approach on how to derive a coherent ethical system, we have set the rules of the game. In the next chapter, it is time to look at what such a system of ethical equality might look like, what ethical principles of equality it might contain. We leave the area of meta-ethics and enter the realm of normative ethics.

Part 2 Theories of equality

In order to understand the ethics of animal equality, we first need to have a very clear picture of the notion of equality. Part 2 of this dissertation is dedicated to a derivation of three different, material principles of equality. These principles are derived from normative ethical systems (contractualism, consequentialism, ethics of care, virtue ethics and deontological ethics). When applying those three principles of equality to sentient beings (animals), we will get a nuanced and clear picture of the theory of animal equality. This move from normative ethics to the applied ethics of animal equality will be discussed in part 3.

Chapter 4 Impartiality and prioritarian equality¹

4.1 Contractarianism universalized

Let's start from the position of a rational egoist. It would not be wise to simply pursue your own benefits all the time, because you might get into trouble with other people. In particular, it might be better to come to an agreement with those other people, not to harm each other and to help each other in need. This is especially true in 'collective action problems': situations wherein different people would all benefit from collaboration with each other, but there is a tendency to cheat, because the cheater would be even better off. However, if everyone started cheating instead of collaborating, then all will be worse off.

Collective action problems are omnipresent (the prisoner's dilemma is a famous example where two suspects have to decide whether to collaborate with each other or not, see section 7.2). As an example, consider picnicking on the beach. If everyone throws away their waste on the beach, no-one will be able to enjoy a beautiful clean beach. So it is better if everyone collaborates by throwing their garbage in the dustbin. But for you, it would be even better if everyone else collaborates, except you: you can enjoy a clean beach but you would not have to make the effort to go to the dustbin. Your best option is to throw away your garbage on the beach. But if everyone started thinking like that, we end up with no collaboration and a dirty beach.

The political philosopher Thomas Hobbes (1588-1679) based his social contract theory on the assumption that people tend to be rationally selfish. This contractarianism can solve collective action problems: everyone is bound by a social contract to collaborate, help each other and not harm each other. Rationally selfish

¹¹ Sections 4.2 and 4.6.8 are based on *Utility and Personal Identity. A Connectivist Approach*, submitted to Philosophical Studies. Section 4.6 is based on *Aversions Behind the Veil of Ignorance*, submitted to Social Choice & Welfare.

people can benefit from mutual advantages of cooperation. But this contract only applies to people having two conditions: 1) an equal power position (e.g. equal bargaining power) and 2) rationality. The equal power position simply comes from the fact that a rational egoist does not have to maintain a social contract with persons in a weaker position (who have no power to harm or help you or have protectors who can harm or help you), because it would be easier to simply exploit them. The rationality condition is obvious: we can only make agreements with people with whom we can negotiate, so those people have to be able to negotiate.

In this contractarianism of the egoist, the moral community (the set of moral patients; those beings who are given moral status) exists of all rational people who have roughly equal power. The latter condition however does not correspond with our moral intuitions. What about those in the weaker positions? What about impartiality? John Rawls (1971) developed a contractualist theory by deleting this equal power condition. In his theory, the moral community is extended to all reasonable and rational beings (in particular all beings who have a sense of justice and a conception of the good). These are beings who are able to perform a thought experiment, which goes under the name of the veil of ignorance. Suppose we are 'impartial observers' sitting behind a veil of ignorance: we have to imagine we will be born as someone on earth, but in order to guarantee impartiality, we don't know yet who we will be. We could be gay, disabled, female, black,... The only thing we can do is derive moral/political laws (respecting known physical laws) that we would prefer in an ideal world where we don't know who we will be.

According to Rawls, those laws will only be applicable to rational beings able to do the thought experiment. As 'impartial observers' behind a veil of ignorance, those rational beings need to have a sense of justice (impartiality) and a rational will. The problem is that this excludes mentally disabled people, babies with a short lifespan and non-human animals. We intuitively see that those beings have a moral status as well: we are not allowed to use mentally disabled persons as merely means, harm animals without good reasons or torture dogs. The condition that only rational beings belong to the moral community is still too partial and it violates moral emotions such as empathy. It is a bit arbitrary to give the veil of ignorance some half thickness: we don't know who we will be, but we know we will be rational agents.

As Rawls proposed an extension of a Hobbesian contractarianism by deleting the equal power condition, I propose a second extension, by deleting the rationality condition. This is the most extreme universalization of contractarianism. We have to make the veil of ignorance as thick as it possibly can be. No criteria are left out: we include all physical entities (in the broadest sense) in the thought experiment. Hence, we could become an electron, a planet, a car, a computer program, an ink stain, a tree, a pig, a person in the year 3000, or whatever. We can now decide what the moral/political laws should look like. We remark that electrons, trees, stains and other objects are likely

not sentient beings. That means that no-one can influence the well-being of a non-sentient being. So if you were a non-sentient object, it doesn't matter to you what happens to you. You would not be aware of anything, you would not like or dislike anything, you would not want anything, you would not experience harm, you would not have preferences, you would not have interests that you care about, your well-being would remain constant at level zero (i.e. it is absent).

The thickening of the veil of ignorance and the extension of contractarianism to all sentient beings are no new ideas, but were already discussed by Van de Veer (1979), Rowlands (1998), Nussbaum, (2006) and Van den Berg (2011). According to Rowlands, who we are (human, pig,...) is just a matter of luck. We did not have a responsibility or choice in this, so we should not be rewarded for being a human. As being human is beyond the control of an individual, it should be judged morally arbitrary.

It is clear that taking this thickest veil of ignorance is the most impartial and least arbitrary thing to do. We automatically come to the criterion of sentience, because sentient beings are the only beings with well-being, and well-being is the only thing that really matters to us behind the veil of ignorance. But what do we mean with well-being and value of life?

4.2 From feelings and well-being to the value of life

The central quantity in our theory of justice is the notion of 'value of life' or 'lifetime well-being'. This is the value that – behind a veil of ignorance – an impartial observer would ascribe to the complete life of a sentient being (i.e. a being that experiences a well-being). The value of life or lifetime well-being is a function of all momentaneous experiences of well-being during a complete life of an individual. Let us analyze this in more detail.

4.2.1 Affective qualia: from experienced feelings to experienced pleasure

Experienced feelings are subjective, private, direct, conscious, qualitative experiences, phenomenological sensations or qualia (Byrne, 2010). These qualia are internal representations that have an attention or focus. For example I consciously feel this book because I can pay attention to the sensation generated through my fingertips (for the connection between qualia and attention, see Ramachandran and Hubbard, 2001). Just before I paid attention to this feeling of touch, I was not aware of it. There was an

unconscious neural activity, and only after I focused on my fingertips, it became a conscious experience or 'quale' of touch. Qualia are often neutral: I don't feel an urge to avoid touching books. When qualia become affective in nature, i.e. when they are evaluated as being positive or negative (when they generate a positive or negative attitude in the individual holding the qualia), they become positive or negative feelings, i.e. pleasure and pain. A needle in my finger generates a quale that I wish to avoid. This quale is called pain and it generates an urge in me to withdraw.

4.2.2 The importance of preferences: from experienced pleasure to momentaneous well-being

Well-being experienced at a specific moment should be distinguished from mere pleasure. I define momentaneous well-being as the composition of all the positive (minus negative) feelings and emotions that are the consequence of (dis)satisfaction of preferences (the things that one wants)². This is an important definition. Its formulation in terms of feelings and preferences combines a mental state account (having mental states such as feelings of pleasure) and a preference satisfaction account of well-being (see Shaw 1999, chapter 2).³

There is a connection between feelings (of pleasure) and preferences (or needs): feelings are nothing but indicators to see when something is met or unmet. This 'something' is a preference or need.⁴ According to the psychology of Maslow (1943),

² These preferences can include some unconscious preferences, in particular dispositional and instrumental preferences. For example, during my sleep, I have unconscious preferences. Those preferences have the disposition to become conscious when I wake up and think about the preferences. An embryo on the other hand, is unconscious but does not (yet) have such dispositional preferences. An example of an unconscious instrumental preference is the preference that an animal has in staying alive, even if the animal does not have a notion of life, death and its own future. Staying alive is important if the animal wants to satisfy other, conscious preferences. See Visak, 2011, p78.

³ One could restrict the preferences to well-informed, rational preferences in order to avoid preferences that are actually bad for us (for example a preference for drugs or a preference to marry someone whom I erroneously believe to be my perfect match). But this restriction might not be necessary in my account of well-being, because the definition of well-being not only contains the condition that preferences be satisfied, but that those preference satisfactions result in an increase of positive feelings (or a decrease of negative feelings). The satisfaction of misinformed, irrational preferences would not generate more positive feelings.

⁴ Sometimes feelings are unreliable in measuring a need. For example a malfunctioning amygdala might generate an irrational fear, i.e. fear when there is no danger and the need for safety is met. The irrational fear does not correspond with an unmet need for safety, so one might think that according to the definition of well-being that I proposed, those negative feelings of irrational fear do not negatively contribute to well-being. The same goes for pains in phantom limbs: those pains do not correspond with an unmet need for bodily integrity, because the body part is absent. However, if the patient with the malfunctioning amygdala

preferences contain not only physical functioning (food, water, movement, rest, health, safety,...) but also for some individuals: social needs (connection, compassion, acceptance, warmth, contribution,...), play (joy, humor,...), autonomy (freedom, space, independence, spontaneity,...), honesty (authenticity, integrity, trust...), peace (equality, harmony, order, beauty,...) and meaning (learning, growth, challenge, efficiency, clarity, creativity, purpose,...).

Needs can have different intensities (e.g. a little hunger vs. being very hungry) and satisfactions can also have different levels (e.g. having access to a little bit vs. a lot of food). The higher the level of satisfaction and the higher the intensity (subjective importance) of the corresponding need, the stronger the positive feeling and the higher the momentaneous well-being will be.

As momentaneous well-being does not look merely at positive and negative feelings such as pleasure and pain, but is restricted to those feelings that are the consequence of preference (dis)satisfaction, we avoid a hedonist position (a mental state account) that only looks at pain and pleasure. The hedonist encounters the problems of the 'experience machine' (Nozick, 1974). Suppose we have an experience machine that can give you feelings of pleasure for the rest of your life, by plugging your brains into this machine. However, the experiences in this machine are related to a world that is not real, and you might have a strong need for authenticity (or connection with reality) that will not be satisfied by this machine. The positive feelings generated by the machine are not the consequence of preference satisfaction, so they do not contribute to well-being as I have defined it. That is why a lot of people will be reluctant to step into this machine.

The veil of ignorance helps to explain why merely feelings of pleasure are not sufficient in an account of well-being. From behind the veil, you know you will be someone who does not prefer to live a life in an experience machine. Hence, this means that a need will not be satisfied and your value of life will be lower. To take another example: suppose behind a veil you can decide between two situations. In the first situation, you will experience pleasure with your lover, and your lover is faithful. In the second situation, you will experience as much pleasure as in the first situation, but your lover is unfaithful and you will never know this (you believe your lover is loyal, and being loyal is very important to you). From a pure mental state account, both situations would be equally preferable, because the happiness is equal. However, if I were an impartial observer behind a veil, I would prefer the first situation. That preference reflects a need to be in contact with reality or with the truth, and it implies that merely

does not want to feel this irrational fear, s/he has an unmet need, i.e. a preference for inner peace. Hence, the irrational fear and the phantom pain do lower someone's well-being according to the definition.

feelings of pleasure are not sufficient in the notion of value of life. If an impartial observer behind the veil prefers a situation where s/he will experience a level of pleasure to a situation where s/he will experience the same level of pleasure in a virtual world of an experience machine, the value of life cannot merely depend on a mental state of pleasure. Something else is important, and that something points to a preference or need for e.g. the truth.⁵

4.2.3 The problem of interpersonal comparability: from individual well-being to comparable momentaneous well-being

An individual can measure its momentaneous well-being and the strength of its feelings and preferences.⁶ The big problem is the comparability between different individuals⁷. Feelings are qualia, and hence they are private: they cannot be objectively measured or communicated. Asking whether the well-being of person *i* in situation *X* is equal to the well-being of person *j* in situation *X*, is like asking whether my perception of red is the same as your perception of red. A theory without interpersonal comparison of well-being has a very serious counter-intuitive implication: different Pareto optimal situations cannot be mutually compared. A Pareto optimal situation is a situation in which it is impossible to make any one better off without making at least one individual worse off. Consider a huge income inequality: if the income of the poorest cannot be improved without lowering the position of the richer person (and if the income of the

⁵ Another aspect that comes into play here, is the principle of rule universalism (see 1.2). Suppose I lie to you by saying that *X* is the case whereas in reality *X* is not the case. And suppose that telling the truth (that *X* is not the case) would decrease your positive and increase your negative feelings about *X*. If you will never know that I lied, you will have the same positive feelings about *X* as when *X* was really the case. However, if I am allowed to lie in this situation, then rule universalism implies that you know that everyone in a similar situation as me is allowed to lie. In that case you still do not know that I actually lied, but you do know that I think it is permissible to lie. This might give you an uncomfortable feeling of insecurity, because you have a need for trust that is not met. As a consequence, you do not want a rule that permits everyone to lie in similar situations. Hence, I am not allowed to lie, even if a well concealed lie does not influence your positive feelings about *X*. In other words: a mental state account of well-being combined with rule universalism can avoid some counter-intuitive implications of a simple mental state account that is not combined with rule universalism.

⁶ If choices *A* and *B* are incomparable for an individual, i.e. if that individual is psychologically unable to estimate whether choice *A* gives him/her a higher well-being than choice *B* (e.g. getting a weak emotion of long duration versus a different emotion which is intense but brief), an impartial observer is permitted to make an own estimate. If the impartial observer would choose *A*, then s/he can make that choice for the individual.

⁷ The literature on interpersonal comparison of well-being is vast. See e.g. Elster & Roemer (1991), Hammond (1976), Harsanyi (1955).

richer person cannot increase without a cost for the poorest), we have a Pareto optimal situation that allows a high inequality.

Well-being differs from income in the sense that it cannot be interpersonally compared. One person can compare his own levels of well-being (as he can compare his perception of red with his perception of green), so the best we can get is a Pareto optimal situation of well-being: even if we cannot compare the well-being levels between the persons, it is possible to know that we cannot increase someone's well-being without lowering the well-being of someone else. As Pareto-optimality still allows for serious inequalities, a theory of equality needs to go beyond this Pareto criterion. Without interpersonal comparability, we would not be able to compare for example the harm of death of person A with the harm of a mere pinprick of person B. We need an interpersonal comparability of well-being if we want to avoid such counter-intuitive implications.⁸

A first step to move further beyond merely Pareto efficiency requires a small deviation into some metaphysics (something outside the positive sciences, because well-being cannot be measured from the outside, just as someone's perception of red cannot be measured from the outside). We have to postulate an ideal observer who has an impartial, fully informed point of view. Ideally, this person has experienced almost anything that anyone can experience, having all kinds of preferences that anyone can have, and s/he has a perfect, unbiased memory to compare the levels of well-being during those experiences, having those preferences.

To avoid too much god-like metaphysics in ethics, we quickly have to move more down to earth. We can try to approach the perspective of this ideal observer, when we use as much of our empathy as we can. This is where the veil of ignorance comes into play: we imagine ourselves in the positions of other beings, using our empathy. We do not know the well-being of someone else, but we can measure our empathic well-being: our estimate of the well-being of the other.⁹

Using this empathy, we can see that someone's potential maximum well-being can be higher, the more needs that being has. In other words, a being with more needs can reach a higher well-being (compared to a being with only a few needs) if all of his or her needs are satisfied. The potential minimum well-being can also be lower when a being

⁸ Furthermore, Arrow (1963) demonstrated some impossibility theorems that occur when well-being is not interpersonally comparable. See also Roemer (1996).

⁹ One might object that it is very difficult to empathise with non-human animals or cognitively disabled humans, because they are so different from us in the sense that they lack e.g. self-consciousness or a concept of one's future and death. Still, with enough imagination, we can make best estimates of their well-being. The experiences of those animals could perhaps be compared with our experiences in certain dream states, where we lack full self-consciousness and a concept of our future, but we still feel fear and pain.

has more needs and when all those needs are not met. As a simplified example, suppose we have a being with only one need. The momentaneous well-being arising from that need can vary from e.g. -1 (needs far from being satisfied, so this being rather prefers to die than to experience this negative feeling), to 0 (needs satisfied to some extent, so that for this being it doesn't matter if s/he lives or dies), to $+1$ (needs highly satisfied). A being with two needs (both of the same intensity), however, can have a well-being ranging from -2 to $+2$, if the contributions of the individual needs can be added. This latter addition property is not a necessary condition for our theory we will discuss, and is only meant for didactical purposes. Here we want to address the possibility that different beings can have different potential levels of well-being.

A problem arises: if you and I use our empathy to estimate the well-being of two individuals, we can get different results. Who has the best estimate? We know that we can both be biased in all kinds of ways. To solve this, we can first study our cognitive biases and try to counter them. Cognitive biases that might influence decision making are e.g. duration neglect, the framing effect, the priming effect, negativity bias, optimism bias, selective perception and fading affect bias (see e.g. Pohl, 2004).

Second, we can communicate and try to move to a consensus. Ideally, all moral agents who have empathy can do the veil of ignorance exercise and work towards a consensus to get the best, unbiased empathic well-being. This consensual empathic well-being best approaches the estimates of the hypothetical ideal person.

Third, we can introduce objective measures as approximations to estimate someone's well-being. These objective measures can be used to counter biases. Examples are primary goods (Rawls, 1971), resources (the resourcist position that looks at economical goods that can be distributed (Dworkin (1981))), capabilities (the sufficientarianist position of the capabilities approach which looks at basic functionings that one is free to choose to improve one's flourishing (Nussbaum 1992, 2000; Sen 1992)), measures of desert (the compensationist position of desert-principles of justice which focus on the compensation of virtuous work (Dick, 1975; Lamont 1994; Milne 1986; Sadurski 1985)). Objective quantities like economic resources, income, wealth, health, jobs, compensations, capabilities or happiness surveys are nothing but approximations of well-being: these elements contribute to well-being, but cannot be reduced or set equal to well-being. They should be used as tools to objectively counter someone's biases when performing the thought experiment of the veil of ignorance. Hopefully, in the future, neuroscientific discoveries could make more accurate estimates and comparisons of well-being possible.

One more thing needs to be said about incomparability. Suppose that individual 1 has a well-being at level A, whereas individual 2 can have four levels of well-being:

$B < B' < B'' < B'''$. Suppose B is so low, that everyone agrees that $B < A$. Similarly, suppose that B''' is so high that everyone agrees that $B''' > A$. But B' and A appear to be incomparable, and the same goes for B'' and A .¹⁰ In that case, an impartial observer behind the veil is permitted to choose for example $A = B'$, and hence $A < B''$. Another impartial observer may choose $A = B''$. An analogy with physics, in particular special relativity, might be handy (see Pivato, 2009). The time dimension corresponds with the level of well-being. The space dimension represents different individuals. An individual with well-being A corresponds with a unique point (event) in space-time. Each event in space-time has a future and a past light cone. The inequality of well-being $B < A$ can be interpreted by the claim that space-time event B is in the past of event A , or more exactly: event B lies in the past light cone of event A . If $B''' > A$, then B''' lies in the future light cone of A . But B' and B'' lie outside the future and past light cones of A (although B'' lies in the future light cone of B' because $B'' > B'$). If B' lies outside of the light cones of A , we can always choose a frame of reference whereby B' and A occur simultaneously, i.e. $B' = A$. But we can also take another frame of reference that gives $B'' = A$. This analogy with special relativity clarifies the intransitivity problem: if B' and A are incomparable and may therefore set equal, then also B'' and A may be set equal. Hence one could naively say that $B'' = A = B'$. But we saw that $B'' > B'$.

We have seen that impartial observers are permitted to make a few estimates and choices of their own, such as the choice between incomparable levels of well-being (the choice of frame of reference in special relativistic terms). In section 4.6.11 on democratic impartial preferences of moral agents, we will see how those different choices of the different impartial observers have to be dealt with in a democratic way.

4.2.4 The lifetime perspective: from momentaneous well-being to the value of life

Until now, I focused on empathic estimates of momentaneous well-being. But behind the veil of ignorance, we have to look at the complete lives of individuals, and attach values to those lives, because of two coherent reasons.

First, with a lifetime perspective we can avoid the replaceability problem: painlessly killing someone and replacing him/her by another individual who has the same momentaneous well-being, is not allowed, even when the aggregate of momentaneous well-being would not decrease (see the discussion in section 4.6.4.3).

¹⁰ Nolt (2013) discussed the relevance of this problem in animal ethics.

Second, people are allowed to choose when in their lives they experience pleasure and pain. Intrapersonal (within the same person's life), intertemporal shifts in well-being are permissible. For example it is permissible for me to eat a lot of candy today, even if as a result I get a toothache tomorrow (note that today I don't have a clear permission of my future self to cause this toothache). By eating candy today, I harm my future self, but that is not immoral. At most it is imprudent. In contrast, it would be immoral if I cause you a toothache without your permission. I do not have to consider my future self as a separate person, but I have to consider you as separate. A mere focus on momentaneous well-being will not be able to make a difference between intrapersonal (but intertemporal) and interpersonal distributions of well-being. This difference corresponds with a moral intuition that there is a difference between imprudent and immoral behavior.

The value of life (I often use 'lifetime well-being' as synonym) corresponds with how much we, behind a veil of ignorance, would prefer to live the complete life of that being. As we saw, momentaneous well-being is not interpersonally comparable, but at least it is a quantity that does not involve moral evaluation: it is a descriptive instead of a normative quantity. Integrating someone's momentaneous well-being into a value of life introduces normative elements. The value of life introduces normative elements: the impartial observer weighs the momentaneous experienced well-being and s/he reflects on this well-being from behind a veil of ignorance. This allows for the introduction of elements deemed important by the impartial observer behind the veil of ignorance. If s/he wants to, the impartial observer behind the veil can introduce elements from an 'objective list account' of well-being (see Shaw 1999, chapter 2; Crisp, 2008).

To study lifetime well-being, we first have to tackle the problem of integrating momentaneous well-being over a period of time, say one second. There is a difference between objective versus subjective rate of time (Bostrom & Yudkowski, 2011). A human eye sees roughly 20 frames per second, whereas the eye of a fly can see movement ten times faster than a human eye. It is as if in one second, a fly experiences more. The objective time is 1 second, but the fly has a ten times faster subjective rate of time. As the fly sees more within that one objective second, it is as if one second for the fly corresponds with ten seconds for a human. The fly experiences everything ten times slower. Suppose the same happens with pain: one individual feels 10 pulses of pain per second, another individual feels 100 pulses per second. Then this second individual has experienced more. The time-integrated pain over one second of time is ten times higher for the second individual. Lifetime well-being should take into account someone's subjective rate of time, not the objective rate of time.

The value of life is a function of (consensual, unbiased, empathic) momentaneous well-being of an individual, but it is not a trivial summation or integration of the momentaneous well-being over all moments of a lifetime, from conception to death. For

example Velleman (1991) argued that a life with constantly increasing momentaneous well-being (starting miserable at birth, ending glorious at death) is preferred to a deteriorating life, even if both lives have the same amount of summed momentaneous well-being. As a second example, consider the argument of the long living oyster (Crisp, 2008)¹¹. Which life would you prefer: the life of a normal human being with life expectancy 80 years, or the life of an oyster with a life expectancy you may choose (a trillion years?), but with a very small but positive and constant well-being?¹² In short, the human being has a high momentaneous well-being for a short period of time, the oyster has a low well-being, but summing this low well-being over the very long course of its life, the total (summed) well-being of the oyster can be higher than that of the human. Yet, a lot of people would prefer being born as the human, no matter how long the life expectancy of the oyster may be. This means that these people value the value of life of the human higher than that of the oyster. Why is that? Perhaps because they expect that leading a human life is less boring, and they have a need for variation or psychological growth. These needs cannot be satisfied in the life of the oyster. Perhaps the oyster does not have those needs, but that still means that a human who has these extra needs and who has satisfied those needs, has a higher value of life.

To solve the problem of the long living oyster, the value of life can be expressed as a trade-off between quantity (length of a life) and quality (average momentaneous well-being). When quantity is low and quality high (a very short but very happy life), it becomes important to increase quantity (life expectancy). When quantity is high and quality low (a very long but moderately happy life), it becomes important to increase quality.

A simplification will clarify the kind of trade-off between quantity and quality. Suppose the value of life can be expressed mathematically as something like: $w.R.T/(R+T)$, with w the (consensual, unbiased, empathic) lifetime-averaged momentaneous experienced well-being¹³, T the total lifespan and R a reference length of

¹¹ Parfit (1984, p.161) described a similar argument, comparing two lives: a Century of Ecstasy (high but temporary well-being) versus Drab Eternity (very long positive but low well-being).

¹² As invertebrates, oysters are perhaps not sentient, but for argument's sake, assume that oysters are sentient beings.

¹³ Kahneman (2011) discovered a difference between an experiencing self (evaluating a currently experienced momentaneous well-being) and a remembering self (evaluating the remembered well-being of a past event). Our distinction between the experienced (momentaneous) well-being w and the value of life can be understood in a similar vein. Although the value of life is distinguished from the momentaneous well-being, it does not equal the remembered well-being. The value of life is evaluated by a hypothetical well-informed fully rational person behind a veil of ignorance, the remembered well-being is evaluated by a real, fallible person remembering a past event.

a life.¹⁴ When T is small, the denominator becomes the constant R , and hence an increase in $w.T$ (the well-being integrated over the lifetime) dominates. When T is large, the value of life approaches the average momentaneous well-being w times R , and hence an increase in w dominates.

As I will demonstrate in a later section (4.6.8), the reference time length R can be understood as a (complex) function of the psychological connectedness (Parfit, 1984), or more precisely: the memories and sense of the future of the individual generates a psychological identity over time¹⁵, making it possible to claim that a person at time t_1 is or is not the same person at a later time t_2 . I am to a large degree a different person than the person I used to be at age ten, although I still count as the same individual.

If the hypothetical sentient oyster has a low psychological connectedness, it gets a small value of R . A normal human being has a high connectedness and hence a higher R . That means that for a human individual, it takes a longer time to change the psychological identity to such a degree that s/he becomes a different person than s/he used to be.

This difference of the reference time length R between different individuals has important consequences in deciding who to save. Suppose we have to decide between extending the life of a normal human versus the life of a sentient non-human animal (e.g. the sentient oyster). We see that extending the life of the oyster does not strongly increase his value of life if his reference time length R is small. Even if his life is extended by many years, it does not much contribute to his value of life. Not much value of life is lost by an earlier death of the oyster. For a human however (supposing a constant momentaneous well-being), an increase in life span results in an almost linear increase in value of life.

¹⁴ For reasons I will not explain here, the value of life might read $w.T$ when the average momentaneous well-being w is negative.

¹⁵ Note that a psychological identity over time is basically the property that distinguishes Regan's subjects-of-a-life criterion (Regan, 1983, p.247) from mere sentience. Subjects-of-a-life not only have the properties of sentience ("perception", "an emotional life", "feelings of pain and pleasure", "preferences", "welfare-interests" and "an individual welfare"), but according to Regan they also have a "psychological identity over time", which includes "memory" and "a sense of the future", "including their own future" (p.247). Similarly Singer's preference utilitarianism (Singer, 1993) gives a priority to beings capable of holding preferences towards the future over those beings who are only concerned with their immediate well-being. Persons who are capable of desiring to continue to live as a subject of experience receive a stronger right to live than sentient beings who lack such personal identity over time. The latter are replaceable, according to Singer. The harm of death is dependent on having a sense of the future and a capability of seeing oneself as existing over time. Self-aware persons who see themselves as continuing selves existing over time are less replaceable, according to Singer. Both Singer's preferences account and Regan's subject-of-a-life account can be supported by a theory that values psychological connectedness.

4.2.5 Personal identity and psychological continuity

Value of life is a function of the momentaneous well-being experienced over a lifetime. But what is the lifetime of a single individual? As just mentioned, different stages in the life of an individual might correspond with different persons. In our daily lives, the life of an individual extends from conception to death (or better: from first till last experience), making it easy to determine what is the complete life of the individual. But Parfit (1984) presented some futuristic thought experiments that challenge the notion of personal identity. A person at a specific moment has a mind that is composed of e.g. memories, beliefs, desires and character traits.

But what happens to a person during e.g. teleportation, when mind and body are destroyed at one place and recreated at another? What if a mind can be multiplied in two exact copies, for example when the teleportation fails, the original mind and body are not destroyed but still a second mind and body are created at the other place? What about imperfect copying a mind M into a slightly different mind M'? What about mind swapping: putting mind M1 (that originally belonged to body B1) into body B2 and mind M2 into body B1?¹⁶ What about multiple personalities, two minds in the same body? What about splitting a mind into two minds? What about fusing two minds into one? Or what about gradual changes of minds and bodies into completely different minds and bodies, where mind M1' in body B1' has elements from both M1 and M2 and both B1 and B2?

In those hypothetical cases, it becomes difficult to make distinctions between different individuals. We should abandon the all-or-nothing relationship of personal identity. One radical option would be to consider a continuum of different individuals, one for each momentaneous mind at each moment of time. However, Parfit (1984) pointed out that an individual can be defined or described by a psychological connectedness and continuity between different momentaneous minds at different moments in time (see also McMahan, 2002). Compare it with a rope, composed of different strands, each strand having a different length. There is no strand that extends from one end of the rope to the other, but still the rope has a connectivity in terms of connectedness and continuity. Two points of the rope are connected if there is a strand that runs from the one to the other point. The more such strands between the two points, the higher the connectedness. Two points of the rope are continuously linked if there are intermediate points such that each two neighboring points are connected (even when the two endpoints are not mutually connected; when there is no strand running between the endpoints). The strands of the rope are the analog of properties of

¹⁶ See e.g. Williams' famous thought experiment of torture (Williams, 1970).

the mind (e.g. memories, opinions and character traits). In this analogy, one end of the rope corresponds with the beginning of life, the other with the end. At the end of your life, you might not remember anything from the beginning of your life, but there is a continuity: an interlinked chain of memories shared by intermediate momentaneous minds. The problem of personal identity is similar to the problem of how to define a rope and how to distinguish one rope from another.

How to deal with this problem of personal identity behind the veil of ignorance? Behind the veil, an impartial observer sees the huge set of all minds at all moments. At each moment, a mind has a unique momentaneous well-being. One option is that the impartial observer groups the set of momentaneous minds in subsets, each subset referring to the complete life of one individual. After putting all momentaneous minds in subsets, the impartial observer looks at a subset that now corresponds to the life of one individual (as defined by the impartial observer). This subset is composed of all momentaneous minds of that individual, each element having a momentaneous well-being that the individual will experience. It is this subset that the impartial observer gives a value, the value of life, which corresponds with how much s/he prefers to experience all experiences of all the momentaneous minds of that individual (that subset).

In most familiar cases, this grouping in subsets is easy and is restricted to strong conditions (i.e. not all possible groupings are allowed). But in the Parfitian situations, the grouping becomes complex and to a degree arbitrary. Therefore, in a later intermezzo I will present another approach how an impartial observer behind the veil can solve this problem of personal identity. That new approach will be more suitable (less arbitrary) to deal with futuristic Parfitian situations of e.g. teleportation, mind copying and mind swapping.

All in all, value of life is the totality of everything one prefers from behind the veil of ignorance, in the expectation to live the complete life of an individual over time. It is everything that would matter to you if you were a sentient being, living its complete life. The value of life is a complex function of momentaneous experienced well-being. As mentioned above, this momentaneous experienced well-being is composed of all the feelings that are the result of (dis)satisfaction of preferences. The term 'experienced well-being' has two words, which means it combines a mental state account (the subjective *experiences* to like things) with a preference satisfaction account (the *well-being* in terms of what one wants). All things that one likes and all things that one wants matter to the experienced well-being. The value of life introduces normative elements: a weighting of the momentaneous experienced well-being and a reflection on this well-being from behind a veil of ignorance.

Value of life is very difficult to measure. All we have is our empathy, our scientific knowledge and our imagination. We have to try placing ourselves in the position of others, by using empathy, by imagining that we could be the other person, with all his

or her needs and feelings. The ‘emotional’ method of sampling empathic feelings and the ‘rational’ method of imagination behind the veil of ignorance are rules of thumb to make educated guesses about the order of the values of life of different individuals. Empathy and imagination are virtues to be developed and already allow us to move quite far.

As all sentient beings have subjective experiences of their feelings and needs, all sentient beings have a lifetime well-being or a value of life for themselves. The model we are about to discuss therefore applies to all sentient beings. It should therefore also include mentally disabled humans and non-human animals. We should not restrict this theory of justice to only rational, self-conscious beings. Hence, a person should be interpreted as an individual who has personal experiences. In this interpretation, a person is equivalent to a sentient being.

Note that different sentient beings, such as a frog and a human, might have strongly different levels of lifetime well-being. There are four reasons why a frog might have a much lower lifetime well-being than a normal human. First, frogs likely have less needs and preferences (e.g. less need for accomplishments or relationships) than most humans. Second, the intensity of preference (dis)satisfaction might be lower in frogs: a frog might have less capacity than most humans to experience pain and pleasure, due to a smaller brain with less neurotransmitters and less receptors (see Vallentyne, 2006). Together, these two reasons imply a much lower momentaneous well-being for the frog. Next, frogs have a much shorter lifespan than those of most humans. And as a fourth reason, frogs likely have less psychological connectivity between different life stages.

In summary, two reasons imply a gap between the momentaneous well-being of a human and a frog. In terms of lifetime well-being, the gap is even bigger due to two additional reasons that refer to the lifespan and the psychological connectivity.

Vallentyne (2006) and Holtug (2007) discussed the far reaching implications of an ethic of redistributive (strict) egalitarianism when animals such as frogs are included, due to the vast difference in levels of well-being. Egalitarianism becomes very demanding for humans, because in order to close the gap between frogs and humans, nearly all resources should go to frogs (and many other non-human animals). The theory that I propose has much less demanding consequences for humans due to two reasons.

First, I propose a prioritarian ethic instead of an egalitarian one (prioritarianism was also suggested by Holtug (2007), but as we will see, my prioritarian ethic has some relevant benefits compared to his). Compared to strict egalitarianism, prioritarianism is more considerate to efficiency: the benefit for the worst-off should not come at a cost of much more lifetime well-being of the better-off.

And second, the above four reasons not only indicate that frogs have lower actual levels of lifetime well-being, but also lower potential levels. This potential level is the level of lifetime well-being that an individual would have when all distributable goods

(all means and resources on earth) are distributed to the maximum benefit of this individual. Their lower potential levels imply that, after distributing all resources on earth to a frog, its lifetime well-being will not increase by the same amount as when all resources are distributed to a human. In other words, compared to frogs, humans can be benefited much more by the same amount of resources. Humans are more efficient than frogs in translating means and resources into lifetime well-being. This higher efficiency is relevant in a prioritarian (but not in an egalitarian) ethic.

Compare it with the problem of distributing an amount of water between different glasses. The level of water in a glass represents the actual level of well-being of an individual; the volume of the glass represents the potential level of that individual. A frog is comparable to a small glass: pouring water into a small glass is more difficult than pouring it into a big glass, resulting in more waste for the small glass. The small glass more easily results in a spill (overflow). This waste of water decreases the efficiency of a distribution of water. Of course, one could increase the volumes of the small glasses, just as one could (genetically) enhance frogs to increase their potential lifetime well-being. But then we do not have the same glasses (frogs) anymore.

4.3 The maximin principle

In the previous section, I discussed the notion of lifetime well-being (value of life). However, nothing has yet been said about how to distribute these quantities. The maximin distribution principle is a theory of justice, favored by John Rawls (1971), which can be derived from the thought experiment of the veil of ignorance. The principle says that we should strive for an increase or **maximization** of the lifetime well-being of the beings in the worst-off position (the beings with a **minimal** amount of well-being). The focus is on the lowest values of life, trying to maximize the lowest levels of lifetime well-being.

Maximin can be derived from the veil of ignorance by realizing that you could be the individual in the worst-off position. Keeping that possibility in mind, you may prefer a world (or a moral law) where this lowest level is increased and maximized. That means you would prefer a society where the lowest levels of lifetime well-being are not so low, such that you no longer worry about getting one of these lowest levels. And importantly: inequality of well-being is only allowed if it is at the advantage of the worst-off positions. Other inequalities of well-being that do not match this condition are not accepted.

Let's give an example with numbers. Suppose there are two sentient beings, and we can choose between different situations. In situation 1, sentient being A has a value of

life level 10, B has level 100. So there is a big inequality. However, this situation is preferable to situation 2, where A and B both have a value of life equal to 5. Situation 1 is also better than situation 3 where A has level 5 and B has level 200, because in situation 1 the worst-off being has a level of 10 instead of 5. Note that in situation 3, the total sum of well-being levels is 205, which is higher than 110 of the first situation. Maximin is therefore different than sum-utilitarianism, because it gives absolute priority to the lowest levels.

The reason why someone would prefer the first situation instead of the third, is that from behind the veil of ignorance, not knowing whether s/he will be A or B, s/he does not want to run the risk of becoming the individual with the worst outcome. Hence, if we have to choose between situations 1, 2 and 3 from behind the veil of ignorance, it is a choice between three games of chance. Which game do we prefer to play? In each game we have an equal probability of becoming individual A or B. But if we have risk aversion (Arrow, 1965; Pratt, 1964), we do not prefer situations 2 and 3, because in those situations we know that the worst-off position has level 5, whereas we could have had level 10 in situation 1. People with maximal risk aversion are real pessimists and always think as if they will become the person in the worst-off position. They ask the question: what if I would be the worst-off person? They would prefer situation 1, even if it has a lower expectation value, because in this situation they at least have a well-being of 10. In general, risk aversion is the reluctance to accept a game of chance with an uncertain outcome rather than another game of chance with a more certain, but possibly lower expected outcome.

On the other hand, someone who is risk neutral would take the sum-utilitarian choice by looking at the total expectation value of well-being (the sum of products of probabilities and levels of well-being). In situation 1, the expectation value is $\frac{1}{2} \times 10 + \frac{1}{2} \times 100 = 55$. In situations 2 and 3 we have respectively $\frac{1}{2} \times 5 + \frac{1}{2} \times 5 = 5$ and $\frac{1}{2} \times 5 + \frac{1}{2} \times 200 = 102,5$. The latter has the highest expectation value, so is preferred by the risk neutral sum-utilitarian (see e.g. Harsanyi, 1953).

John Rawls' theory of justice (1971) incorporates the maximin principle. When it comes to animal ethics, Richard Ryder (2001) can be considered as the advocate of the maximin principle. His theory of 'painism' gives an absolute priority to the so called maximum sufferer, the sentient being who is in most pain. This is clearly the worst off position. In his theory of animal rights, Tom Regan (1983) also offered two principles: the miniride principle and the worst-off principle. Regan's two principles can be interpreted in such a way that they can be unified in the one principle of maximin.

The miniride principle says: "Special consideration aside, when we must choose between overriding the rights of many who are innocent or the rights of few who are innocent, and when each affected individual will be harmed in a *prima facie* comparable way, then we ought to choose to override the rights of the few in preference to overriding the rights of the many." (p.305) To take an example, suppose we have to

choose between situation X where one individual would suffer and has a value of life equal to 5, whereas ten others would have well-being at level 10, and situation Y where the first individual has level 10 and the ten others have all level 5. The harm done to each individual is the same (a drop of well-being of 5 levels). The miniride principle prefers situation X, and this is also what maximin would say.

The worst-off principle says that: “Special considerations aside, when we must decide to override the rights of the many or the rights of the few who are innocent, and when the harm faced by the few would make them worse-off than any of the many would be if the other option were chosen, then we ought to override the rights of the many.” (p.308) As an example, in situation X one individual has well-being 2 whereas the other ten have well-being 10. Situation Y is similar to the previous example: the first individual has 10 and the others have 5. The worst-off principle and the maximin principle both say that we have to prefer situation Y, because in situation X, the harm done to the first individual is a drop of 8 levels of well-being. That’s worse off than the other people in situation Y.

The worst-off principle strikes many people as counter-intuitive in some extreme examples. What if instead of harming ten people in situation Y, we harmed a million people? The worst-off principle lacks a kind of efficiency. The quasi-maximin principle that I am going to discuss in the next section, would be more compatible with an intuitive judgment that to some degree efficiency is important in distributing well-being.

4.4 The quasi-maximin principle and prioritarianism

We saw that from behind the veil of ignorance, we could arrive at two different theories of justice, depending on our level of risk aversion. Someone who has maximal risk aversion prefers the maximin strategy. A risk neutral person prefers the sum-utilitarian strategy. These two strategies are but two options in a continuum of theories of justice, because there is a continuum in the level of risk aversion. Most people have a high but not maximal level of risk aversion. So let’s take a look at another example. In situation 1, person A had a well-being of 10, B had level 100. In situation 4, we can increase the well-being of A by a negligible amount to level 10,01. In order to do this, the level of B has to drop a lot, to level 11. It’s as if we drive B to extreme poverty in order to increase the level of the extremely poor person A with a negligible amount. According to maximin, we would prefer situation 4, because 10,01 is higher than 10. However a person with high but not maximal risk aversion would still prefer situation 1. This person would adopt a quasi-maximin principle of justice. It is almost but not completely maximin.

There is another way to arrive at quasi-maximin. Our empathy is directed towards the worst-off individual, which is sentient being A in the above example. But if we have a low but not zero need for efficiency, we would not prefer situation 4. It doesn't seem efficient to drop B in order to increase A with just a tiny amount. It's too much a waste of well-being.

Therefore, we have two reasons to prefer situation 1: *impartiality with a high but not maximal level of risk aversion (need for safety), and empathy with a low but not zero need for efficiency*. These two reasons cohere with each other and they are both based on moral intuitions of impartiality, safety, empathy and efficiency. The two reasons correspond with a rational and an emotional approach, and with two viewpoints: the rational approach looks at a situation from the outside, from an impartial point of view behind a veil of ignorance. The emotional approach is more down to earth: it looks at a situation from the inside, from the subjective experience of compassion with others. These two coherent approaches give us some justification for a quasi-maximin principle of justice.¹⁷

The quasi-maximin (QMM) principle for a just distribution of values of life.

Maximize the values of life (lifetime well-being levels) of all sentient beings, giving a strong priority on increasing the lowest values of life. I.e. maximize the values of life of the worst off individuals, unless this is at the expense of much more well-being of others.

This QMM-principle gives a high but not maximum priority to the worst-off individuals. It is therefore a kind of prioritarianism. In prioritarianism, the well-being of an individual is weighted with a priority function. The lower someone's well-being, the higher his/her priority. As sum-utilitarianism maximizes the sum of well-being levels, prioritarianism maximizes the sum of weighted well-being levels. I refer to the mathematical section below for more details. But first, let's discuss some applications of this QMM-theory.

¹⁷ For some moral agents, these two approaches might be different. For example someone who has zero risk aversion but a high empathic concern for the worst-off and a low need for efficiency will get two different ethics. According to the impartial veil of ignorance approach, this moral agent would be a sum-utilitarian. According to his/her moral intuitions of empathy and efficiency, s/he would be more maximin-prioritarian. If such a dichotomy occurs, the moral agent is allowed to take his/her preferred approach to determine the level of priority for the worst-off (this level is then democratically averaged together with the preferences of all other moral agents, as will be discussed in section 4.6.11). Furthermore, I expect that most moral agents have some (non-zero) level of risk aversion, and most moral agents have a non-absolute need for efficiency (a non-zero priority for the worst-off). So even when both approaches might differ for one moral agents, when we look at the group of all moral agents, we can expect that they might still easily reach a rather big consensus on the non-zero level of priority for the worst-off. Looking at an individual, both approaches might be mutually incoherent, but on the level of the whole group, they might still be more coherent with each other.

4.5 Applications of the quasi-maximin theory

4.5.1 Rawls' theory of justice

Although the comparison between values of life of different individuals in different situations is very difficult, we can derive a set of approximate rules of thumb that can move us closer to the QMM-distribution of value of life. In his theory of justice, John Rawls derived three such principles (Rawls, 1971, 2001):

1) Equality of basic liberties and rights.

2) Equality of fair opportunity: if individuals have the same ambition and native talents relevant for a position that generates a benefit (e.g. a job), they should have the same prospects of success in competition for that position (see also Arneson, 2008).

3) Equality of economic goods in terms of the difference principle: the distribution of economic goods should be according to maximin. That means that economic inequalities should be in the greatest benefit of the least advantaged persons.

These Rawlsian equality principles can be considered as rules of thumb to approach a QMM-distribution of well-being. Let's first look at equality of basic liberties and rights. We only have to consider rights and liberties that clearly affect the value of life. Take for example the right to free speech. If I have a need for sharing ideas, I will feel frustrated when I do not have the right to free speech, and this obstruction will lower my value of life. However, there are some speech acts (e.g. hate speech or insults) that can lower the value of life of other people (the receivers). In most cases, allowing these disdainful speech acts will violate the QMM-principle. First, as Rosenberg notes (Rosenberg, 2003), someone uttering disdainful speech acts often implies that this person has unmet needs. Insults are a tragic expression of a person with an unmet need. If your boss insults you by saying that you are lazy, this most likely means that your boss feels frustrated and has an unmet need for e.g. efficiency, and that he only found a tragic way to express himself. Also hate speech and scapegoats indicate some unmet need (e.g. for social security or respect).

Let's try to apply our QMM-model to this problem. As a starting point, we have two persons. In situation X, there is no free speech. By lack of further details, and by the symmetry between the persons, we have to assume that a priori (all else equal) both persons have equal value of life, say level 100. This equality is an important assumption in dealing with these kinds of problems. In situation Y, there is free speech, and as a consequence, person B insults person A. The value of life of person B increases to 101, but for person A it decreases to 99. Situation Y violates the QMM-principle. To summarize: not all speech acts satisfy the QMM-principle.

Moving to the second Rawlsian principle, where there is a scarcity of social, economic or political positions (education, jobs, elections,...), the equality of fair opportunity (and

participation) can be derived from the original position. Only someone who is more talented, motivated, trustworthy or experienced to do a job that is beneficial to the least advantaged persons (or more generally a socially beneficial job that helps approaching the QMM-distribution of lifetime well-being), should have a higher prospect to get that job. Hence, equality of fair opportunity is a derivative of the QMM-principle.

The third Rawlsian principle (the difference principle) can also be easily restated in the QMM-framework. First note that this latter Rawlsian difference principle refers to economic goods and not to the value of life (the lifetime well-being). Economic goods (income, resources, wealth,...) only constitute a subset of factors that contribute to the value of life. The QMM-theory as described in this section is more in line with the welfare based principles (like utilitarianism), and hence also incorporates the distribution of liberties, opportunities, capabilities and all other factors that contribute to the value of life.

Ideally, the economic goods should be distributed according to the rule that realizes a quasi-maximin distribution of lifetime well-being. This means that for example disabled persons should get relatively more economic goods in order to compensate for their loss of well-being, except when the transfer of economic goods to these disabled persons cannot be done in a sufficiently efficient way. In other words, when we are only capable of increasing the well-being of the disabled person by a negligible amount by transferring huge amounts of resources to these disabled persons, we should not opt for the transfer.

But as it is often difficult to determine the optimal distribution of economic goods, the economic goods can more easily be distributed according to Rawls' difference principle. So Rawls' difference principle can be considered as an approximation of the QMM-theory.

4.5.2 Responsibility and desert

So far for Rawls' difference principle. Let us also take a look at resource-based (or responsibility-based) and desert-based principles.

In the resource-based principles of justice (Dworkin, 1981), one is concerned about the importance of personal responsibility. According to the QMM-theory, society should not keep on pouring resources down the drain, if worse-off people act very irresponsibly with these given resources (when they negligently squander them) or if they are highly inefficient in transforming these resources into lifetime well-being (see Cohen 1989, Arneson 1989, Roemer 1996).

Consider first the issue of acting irresponsibly with given resources. Some facts that influence well-being (e.g. being born with talents or discovering new resources by brute luck) are beyond someone's control or responsibility. But, given an amount of resources,

an individual has a personal choice and hence a personal responsibility to transfer these resources into well-being. What if s/he makes imprudent or irrational choices that squander resources? Or what if in the hospital we have to choose between helping two patients who are equally bad off and who can be equally benefited by a medical operation; the first one has a genetic disease, the second had a car accident because she was a reckless driver? Luck consequentialism (or responsibility-sensitive consequentialism) claims that the part of someone's well-being that is under responsibility of the individual should not matter in calculating the best distribution of well-being. So how much should responsibility and brute luck play a role in the distribution of well-being? Let me make three remarks on this.

First, it might be likely that there is no such thing as a free will. People might make bad choices (e.g. reckless driving, gambling or being addicted), but they are not responsible for choosing brains that make them vulnerable for those bad choices, just as persons with genetic diseases are not responsible for choosing the bad genes. Hence, we might overestimate the role of personal responsibility. As having a certain brain is a result of brute luck, a lot (or all?) of our personal choices might in the end be beyond our control, beyond our responsibility.

Second, in the hospital example (choosing between the reckless driver and the person with the genetic disease), the choice who to help might influence the distribution of well-being. A choice to help reckless drivers (or other people who make bad choices) might give wrong incentives to some people. For example people might become less dissuaded to make some bad choices. In this sense, personal responsibility plays only an instrumental role. Similarly, when lazy workers or imprudent people experience a disadvantage due to their choice to be lazy or imprudent, they should be helped, benefited or rewarded less compared to the hard working and prudent people. The level of benefits and rewards should be tuned to give the optimal incentives for everyone to reach a QMM-distribution of well-being.

Third, from a lifetime perspective, we have to take into account that the reckless driver, the lazy worker or the drug addict already enjoyed a benefit in the past (the pleasure of driving recklessly, the pleasure of relaxing at work, the pleasure of the drugs). This benefit in the past means that they have less right to a benefit in the future, compared to someone who had brute bad luck (all else equal).

As a result of these three considerations, I am tempted to minimize the importance of the distinction between well-being that is a result of brute luck and well-being that is a result of personal choices. All types of well-being are important in QMM-theory, and the difference between brute luck and personal choice can only play an instrumental role in tuning incentives for behavior.

Now consider the second issue, the problem that people might be highly inefficient in transforming resources into value of life. QMM-theory keeps track of the inefficiencies when distributing benefits. For example consider a benefit that generates a well-being

of 5 units to a well-off person who has initial well-being 5. If this person gets the benefit, his well-being will end up at the level 10. Now consider a redistribution of this benefit from the well-off person to a worse-off person having initial well-being 1. As the worse-off person is less efficient in transforming the benefit into well-being, she will only receive an extra 3 units of well-being, ending up at level 4. Hence we have to decide between option $X=(10;1)$ and option $Y=(5;4)$. QMM-theory prefers option Y.

We can distinguish between two kinds of inefficiencies in the transformation of resources into lifetime well-being. First, there are the things that are beyond the control of the individual: an individual might have medical needs such that a lot of resources are required to generate a unit of well-being. The second inefficiency occurs in the development of e.g. expensive tastes. Having an expensive taste means that one needs a lot of resources to satisfy the taste and to increase the well-being with one unit. Here we should make a distinction between modifiability and satisfiability of preferences. Modifiability means that an individual can influence the presence of the preference: the individual has some power to switch the preference on or off. This modifiability should be distinguished from satisfiability: the power of an individual to satisfy a preference.

Expensive tastes are not only inefficient, they are modifiable, and this property of modifiability is highly morally relevant. According to our QMM-theory, as people are responsible for developing expensive tastes, they have a duty not to develop those modifiable tastes, because those tastes generate extra inefficiencies. Instead of pouring resources down the drain, therapy (e.g. meditation) can be a cheap method to conquer those modifiable expensive tastes and addictions. And in order to dissuade people to develop expensive tastes, we should refrain from redistributing resources to satisfy expensive tastes.

Next to responsibility is the issue of desert. In the desert-based principles, one wants to emphasize effort (Sadurski 1985, Milne 1986) or costs incurred in work (Dick 1975, Lamont 1994), or someone's contribution to society (Miller 1976, Riley 1989). Hence, the notion of desert that is used in QMM-theory is based on two aspects. First, it has a compensationist approach: compensate for the efforts, costs or risks taken by an agent's past actions. Second, it can refer to virtuous actions that contribute to the well-being of others.

Compensation. According to the desert principle, we should distribute economic goods corresponding to the virtue or deservingness of a person (see e.g. Kagan 1999). According to desert-based principles, things done in the past (e.g. someone who worked hard yesterday) influence the just distribution of current resources (e.g. higher payment for the one who worked hard yesterday). The QMM-theory uses a lifetime perspective (a focus on lifetime well-being), and this lifetime perspective allows to take into account an agent's past actions. Hence, the lifetime approach allows for a compensationist desert-based principle. Hard work in the past means that someone's momentaneous well-being in the past is low. This low past momentaneous well-being can be

compensated by a higher future momentaneous well-being to increase the lifetime well-being.

Contribution. We can interpret virtuous work as work that contributes to the society, and more specifically that promotes the QMM-distribution of well-being. The more someone contributes to QMM, the more she should be rewarded in order to support her choice for QMM. And the more her value of life decreases by doing this important work (e.g. by doing hard, long, boring or dangerous work), the more she should be compensated for her loss of well-being. So the more her value of life decreases and the more her work contributes to QMM, the more virtuous and deserving she is.

Someone who contributes more to the well-being of the worst-off persons, should get prior access to more economic goods. For example a nurse should receive a higher income than a professional athlete, because the nurse's contribution to the value of life of the worst-off individuals is higher. Free market distributions of economic wealth are not always compatible with the QMM-theory.

In a desert-based theory of justice, one often adds the 'greater gap principle'. The greater the gap between what someone deserves and what someone has, the more priority should be given for decreasing that gap. In a sense, this is a generalization of prioritarianism as defined above, where priority should be given to the most deserving person. The more deserving person is not always the worst-off person, but can also be the more virtuous person. So not only the well-being of an individual should matter (as in simple prioritarianism), but also someone's contribution to society (to approach the QMM-distribution, i.e. to contribute to the total weighted well-being of all people in society) should be rewarded. And this latter reward is only possible when it is not in conflict with the QMM-theory itself.

Consider the following example. Suppose we have an ill person (well-being level 1), a poor physician (level 10) and a very rich and wealthy person (level 100). The value of life distribution in situation X can be described with the three values (1;10;100). Now, the rich person can give money to the physician so that the physician is motivated to heal the ill person. We now get situation Y=(10;30;30). Situation Y is better than X according to QMM. Note that the increase in well-being of the physician ($30 - 10 = 20$) can be larger than the increase of well-being of the ill person (which equals 9 in the example). We might compare this desert principle with a negative feedback mechanism which acts as a stable attractor when there are disturbances that push us away from the QMM-equilibrium position. This feedback mechanism actively pulls us back towards the state of QMM, by stimulating (rewarding) people who contribute most to the total priority weighted well-being of society.

4.5.3 Future orientation and restorative justice

In the above section we saw that responsibility plays two roles in QMM-theory. First, the lifetime perspective of QMM-theory allows us to take past actions into account, in a way that some compensationist notion of desert becomes important. Second, the QMM-theory also looks at how we can praise or blame people to influence their future behavior. Hence, the QMM-theory has both a past and forward looking aspect.

The forward looking aspect has major implications for the criminal justice system that needs to be revised. If the behavior of a person is in strong violation of the QMM-principle (and with other ethical principles to be discussed in next chapters), then some rights of that person should be taken away in order to protect society from future violations of the QMM-principle. Especially when we know that someone has malicious intentions, it is likely that this person will violate the QMM-principle in the future. Imprisonment should be considered as a kind of quarantine to protect society from threats (such as murderers, pathogens) that endanger a QMM-distribution of lifetime well-being. Moral responsibility for criminal behavior (i.e. behavior that deviates from the QMM-principle) should be a measure of the likelihood that the person will do other crimes in the future (because e.g. his/her brains are wired in a certain way that makes him/her more susceptible to do crimes). The probability of recidivism (i.e. the risk that someone might perform actions in the future that deviate from the QMM-principle) should be taken into account when liberties and rights are distributed.¹⁸

Some behavior such as stealing or lying would be permitted, however, if the behavior is in correspondence with the QMM-principle (for example a poor thief who steals from the rich, a person who lies to protect someone's life).

Restorative justice might be preferred to retributive justice, because it might be better for the well-being of both victims and perpetrators. Most perpetrators typically are victims themselves who tend to have a low well-being (they might have strong

¹⁸ For example my right to use self-defense against someone who is about to harm me depends on the moral responsibility of that person who is a threat for me. The more responsible s/he is, the more likely s/he will be a threat to others in the future. Consider a) a construction worker who stumbles by accident and falls right on me, b) a reckless car driver who is about to hit me by accident with the car and c) a murderer who intends to kill me. In all three cases, I can defend myself, by doing something that will harm the threat (the worker, driver or murderer). But the level of harm that I am allowed to use depends on the level of moral responsibility of the threat. The more responsible, the more harm I am allowed to cause to the threat in defending myself (still avoiding any unnecessary harm). This moral responsibility measures the likelihood that the threat will be a threat again in the future. It is more likely that a murderer will commit a similar crime in the future than that a car driver will again hit someone, and the latter is still more likely than the probability of the construction worker falling again by accident right on someone else. As a consequence, I am more entitled to kill the murderer in self-defense, than to kill the construction worker.

feelings of frustration due to discrimination, lack of education, lack of opportunities, or traumatic experiences in the past). The QMM-principle, combined with neuroscientific evidence (about e.g. the lack of free will), implies that the legal justice system should be more forward looking (a restorative justice that focuses on how to improve well-being and how to most efficiently change a criminal's brain and behavior) instead of backward looking (a retributive justice that focuses on punishment, guilt and blame). The only backward looking part in QMM-theory has to do with compensations for past actions, which relates to a notion of desert as we saw in the previous section. Instead of punishing people as retributive justice, it is better to create circumstances in such a way that people tend to behave more morally.

To summarize, we see that the QMM-theory combines and encompasses a lot of different ideas: prioritarianism (keeping the balance between Rawlsian maximin and sum-utilitarianism) and desert-based, welfare-based and resource-based theories.

4.6 Aversions behind the veil of ignorance: a mathematical description for a theory of justice

4.6.1 Why a mathematical model?

In this section I want to unify different theories of justice and equality, by placing them in a coherent framework. In order to do this, I will try to use mathematical modeling as much as possible. Economists and natural scientists are familiar with the use of mathematical models. In moral philosophy however, only a few theories of justice (e.g. utilitarianism) have some more or less explicit reference to quantitative objects (e.g. utility).

Using a mathematical framework will help us to see different theories of justice and their mutual relationships in more clarity. Mathematical modeling offers an efficient toolbox that helps us to work towards a more unified theory of justice. The mathematical equations in this section should therefore not to be taken too literally, but they should be used as ways to simplify expressions of complex ideas. What I will attempt to do, is combine different theories (utilitarianism, maximin, prioritarianism and egalitarianism) into a mathematical expression that contains some parameters. These parameters can take different values, and for specific values we get a specific theory of justice.

As mentioned in a previous section, there are two arguments for quasi-maximin (QMM) prioritarianism: one is based on a Rawlsian argument of impartiality (the veil of

ignorance), whereby we assume that the person in the original position has a high but not absolute need for safety (high but not maximum risk aversion), and one based on empathy for the worst-off individuals, combined with a low but non-zero need for efficiency in terms of well-being. Hence, efficiency is inversely related to risk aversion. In this section I derive a mathematical formulation of the quasi-maximin prioritarian principle, using the veil of ignorance as starting point.

4.6.2 The mathematics of consequentialist welfare ethics

Prioritarianism is a consequentialist theory that looks at the outcomes of actions in terms of well-being. It was made popular by Parfit (1991, 1997) and states that we should maximize everyone's well-being, giving priority to the worst-off individuals. As a consequentialist theory, it lends itself to mathematical modeling using e.g. utility functions. In particular a priority weighted utility function is used to describe prioritarianism (see e.g. Broome 1991; Brown, 2007; Holtug 2006; Rabinowicz 2002; McCarthy 2003, 2008). These utility functions are the elements of a welfare function, a quantity that represents the consequentialist betterness relations between different choices (different situations or world histories).

A consequentialist welfare ethic such as prioritarianism faces serious problems when it comes to choices involving variable and future populations. These problems are relevant in animal ethics, because animals are consciously bred and brought into existence by our choices. Population ethics (Arrhenius, 2000; Blackorby et al., 2005) is the branch of ethics that deals with variable populations. Population ethics is perhaps the branch of ethics that is mostly plagued with impossibility theorems: using mathematics, some ethicists proved that we cannot find a theory or welfare function for variable populations that meets certain basic moral intuitions. Always some moral intuition has to fall (for an overview of such impossibility theorems, see e.g. Arrhenius, 2000; Blackorby et al. 2003). The goal is therefore reduced to finding a welfare function that still satisfies the strongest moral intuitions regarding variable populations, such that only the weakest intuitions are violated.

In this section I derive a general welfare function from a 'veil of ignorance' thought experiment (Harsanyi, 1953; Rawls, 1971), borrowing some concepts of prospect theory (Kahneman & Tversky, 1979). In particular, I suppose that the impartial observer (decision maker) behind the veil can have different decision aversions: risk aversion (Arrow, 1965), loss aversion (Kahneman & Tversky, 1984) and uncertainty aversion (Epstein, 1999). Rawls (1971) took only the latter uncertainty aversion to arrive at his maximin principle behind a veil of ignorance, but an impartial observer might have or use other aversions as well. Those aversions set the parameters in the welfare function. Applying the veil of ignorance to population ethical situations (problems with variable

populations and potential beings), is tricky. But as I will demonstrate, the welfare function that corresponds with those three aversions also corresponds with some moral intuitions in population ethics.

Risk aversion deals with the problem that the impartial observer behind the veil does not know whose life s/he will live once the veil is lifted. The observer has a probability to become any of the individuals born in the real world, with known, uniform probability distribution: s/he has a probability $1/N$ to become any of the N individuals. Having risk aversion results in a prioritarian ethic: it corresponds with the moral intuition that some priority for the worst-off is important. This prioritarian intuition reflects a trade-off between efficiency and equality. Combining prioritarianism with a lifetime perspective, where the lifetime well-being levels count as the utility variables in the welfare function, also solves the replaceability and non-identity problems (Parfit, 1984). Hence, risk aversion behind a veil of ignorance results in a welfare function that is consistent with strong moral intuitions about efficiency, equality and replaceability.

Loss aversion deals with an asymmetry between preferences for gains and losses. People have a tendency to prefer avoiding losses to acquiring gains (Kahneman & Tversky, 1984). An impartial observer with loss aversion can fix the parameters of the welfare function such that it includes number-dampening population factors. These population factors allow avoiding some counter-intuitive conclusions in population ethics: the repugnant conclusion (Parfit, 1984; Arrhenius et al., 2010) and the reverse repugnant conclusion for positive levels of well-being and the strong sadistic conclusion for negative levels of well-being (Arrhenius, 2000). Hence, loss aversion behind a veil of ignorance results in a welfare function that is consistent with strong moral intuitions about variable populations.

Uncertainty aversion occurs when the veil is thickened in a way that the impartial observer no longer knows the probability to become any of the individuals. It reflects a preference for known risks over unknown risks: when the possible outcomes and the probability to become an individual are known, the risks are known. When the veil is thickened, the risks are not known. This lack of knowledge of risks, which is stronger than the lack of knowledge of outcomes, influences the preferences of the impartial observer. I will argue below that this uncertainty aversion generates a second kind of prioritarian theory called moderate egalitarianism (Jensen, 2003). It differs from prioritarianism in the sense that the level of priority for the worst-off does not depend on the absolute values of lifetime well-being of the worst-off, as in prioritarianism, but depends on the relative positions of the worst-off, relative to the better-off. Moderate egalitarianism has a generalized Gini welfare function (Weymark, 1981). This moderate egalitarianism solves the intransitivity problem (Temkin, 1987) and the problem of the misery for the ultra rich (Dorsey, 2009), at the serious cost of losing independence (or strong separability; see McCarthy, 2008). The problem of independence is related to Allais paradox (Allais, 1953).

I will demonstrate that combining the three aversions, together with the reflection effect of prospect theory (Kahneman & Tversky, 1979), results in a welfare function as a sum of two terms: a positive, number-dampened, weighted power mean prioritarianism, and a negative, weighted total utilitarianism. Including the number-dampening factor in the first term generates a trade-off between quantity (the population size), quality (efficiency in terms of maximally increasing everyone's well-being) and equality (equalizing well-being).

The weighted power mean of the first term contains free parameters. The power p of the power mean can vary from minus infinity, which results in a maximin theory, to 1, which results in a weighted average version of moderate egalitarianism. A negative value for this power corresponds with a quasi-maximin (QMM) prioritarian theory. The weight factors in the power mean can also take different values, ranging from an absolute weight for the worst-off individual, which corresponds with maximin, to a uniform distribution of weights, resulting in unweighted power mean prioritarianism. When the power is 1 and the weights are uniform, we get average sum-utilitarianism. Hence, there are two ways to move from sum-utilitarianism to maximin: using a power mean and using a weighted averaging. These two ways are based on respectively risk aversion and uncertainty aversion.

However, some mathematically proven impossibility theorems in the literature (see e.g. Arrhenius, 2000; Blackorby et al. 2003) indicate that the proposed welfare function violates some moral intuitions in population ethics. I discuss the three most important counter-intuitive implications of the number-dampened prioritarian theory. These moral intuitions might be moral illusions, and I briefly present solutions or ways to deal with those two problems.

Another challenge for prioritarianism, apart from the problems related to variable populations, are lotteries. A lottery represents a policy choice which has different possible world histories as outcomes. A probability distribution over world histories introduces some new complications.

Finally, I will re-examine the problem of interpersonal comparability of lifetime well-being. Each moral agent can perform the thought experiment of the veil of ignorance, using his own risk attitude and his own evaluations of lifetime well-being. To make the theory as objective as possible, we look at distributions of measurable distributable goods (i.e. resources and liberties distributed among all sentient beings). Each moral agent can maximize his own preferred welfare function, respecting the constraints on those distributable goods. Hence, each moral agent can derive his own optimal distribution of goods, and based on a democratic equality of all moral agents, we can take an average of all those distributions as the impartial, optimal distribution of goods.

4.6.3 The impartial observer behind the veil of ignorance

Imagine there are N_B potential beings behind a veil of ignorance. You are one of them. When the veil is lifted, you will live the life of a sentient being in the real world, but behind the veil you do not know yet who you will be. You are an impartial decision maker (an impartial observer) behind the veil, and you can decide between different world histories. In order to study world histories, you can first look at a finite time interval Δt . In this time interval, the number of sentient beings born in front of the veil (i.e. born in the real world) is finite. You can be born and live the life of one of these sentient beings. After deriving the optimal world history that you prefer for this finite time interval, you can take a longer time interval and perform the same process. If the time interval gets longer, more future beings are taken into account, and you might derive a slightly different optimal world history for that longer time interval. In theory, this process should be repeated to the limit of an infinite time interval, encompassing the complete future containing a potential infinite number of beings. But in practice, it will be enough to stop at a sufficiently long time interval.

So consider a world history h limited to the time interval Δt . In this world history, a number $N_F(h)$ of individuals are born in *front* of the veil during that time interval. The number of beings behind the veil, N_B , is equal to or larger than N_F . The difference between N_B and N_F is the number N_U of unborn beings, the potential beings who are not born in the relevant time interval of the world history. The population size N_F can be split in three parts: N_+ is the number of individuals with a positive lifetime well-being, N_0 the number with a zero well-being and N_- the number with a negative well-being.

Each individual i in world history h has a lifetime well-being $x_i(h)$.¹⁹ For all $i \in [1, N_F]$ we can define

¹⁹ We have to assume some properties for the lifetime well-being levels. First, the values x_i and y_i for the same person i are ordinal numbers, which means they can be ordered in a complete well-ordered set. In other words, it is meaningful to say that e.g. $x_i > y_i$, even though these values cannot be quantified. The order relation is complete if for all x_i and y_i we have either $x_i > y_i$, $x_i < y_i$ or $x_i \approx y_i$. This assumption is not a strong assumption: in nearly all our choices we can compare our different needs and feelings affecting our value of life. We might prefer visiting a friend over reading a book, we might prefer short term satisfaction of one need over long term satisfaction of another need,... So we are able to compare the values of life of different choices.

A much more difficult assumption is the following step: there is an ordinality relationship between different individuals. I.e.: we should be able to compare x_1 with x_2 . This is the central most difficult (or vulnerable) point in our theory of justice: how to compare the value of life of different individuals? Is my satisfaction of visiting a friend higher than your satisfaction of reading a book? There is no clear method to solve these kind of questions. All we have are two heuristic methods: empathy and the Rawlsian thought experiment of the veil of ignorance (Rawls, 1971; 2001).

$$x_i^+ = \begin{cases} x_i & \text{if } x_i \geq 0 \\ 0 & \text{if } x_i < 0 \end{cases}$$

i.e. the well-being vector $x^+ = \{x_i^+\}_{i=1}^{N_F}$ is a projection of all well-being levels on \mathbb{R}_+ , the positive and zero values. Similarly, x^- is a projection on all negative and zero values.

4.6.3.1 The welfare function

As an impartial observer, you can construct a welfare function W , i.e. a function that evaluates world histories. The world history which has the highest value should be preferred: if $W(h) > W(g)$, then world history h is better than world history g . We can start with the following general expression for the welfare function:

$$W(x(h)) = P_{s,N_R}(N_+) \langle x^+ \rangle_{a,p} - P_{t,N_R}(N_-) \langle |x^-| \rangle_{b,q},$$

with

$$P_{s,N_R}(N_+) = \frac{N_+}{sN_+ + N_R}$$

a population factor (a number-damping function of N_+ containing parameter $s \in [0,1]$ and a constant reference population size N_R), and

$$\langle x^+ \rangle_{a,p} = \left(\frac{1}{N_F} \sum_{i=1}^{N_F} a_i (x_i^+)^p \right)^{\frac{1}{p}}$$

the weighted power mean (Hölder mean²⁰) of x^+ with power p between $]-\infty, 1]$ and weights a_i such that $\sum_{i=1}^{N_F} a_i = N_F$. The same goes for the negative well-being levels with parameters $t, \{b_i\}_{i=1}^{N_F}$ and q .

A third assumption we have to make seems a big leap into superficiality: the values x_i and y_j are assumed to be cardinal numbers, i.e. quantitative numbers that can be multiplied, added, subtracted,... Although this step might seem superficial, it is in fact only for didactical purposes that I assume cardinality, because now we can use clear mathematical expressions. Therefore, I will speak of a quantitative “model” for a theory of justice.

²⁰ The power mean in the welfare function can be further generalized to a weighted generalized f-mean or weighted Kolmogorov mean (Kolmogorov, 1930) with an invertible function f :

$$W_f(x(h)) = f^{-1} \left(\sum_{i=1}^N a_i f(x_i(h)) \right).$$

4.6.3.2 Invariances

The above general expression of the welfare function satisfies some important invariances.

A first property of the welfare function is time scale invariance. The number of individuals with positive well-being N_+ born in time interval Δt is proportional to the time interval Δt , and the same goes for N_- . We can write the reference population size N_r as $B_r \Delta t$, with B_r a constant reference birth rate. Now all numerators and denominators in the welfare function are linear in Δt , which allows us to take the limit $\Delta t \rightarrow \infty$ without expanding or shrinking a term in the welfare function.

A second invariance is the scale invariance of the well-being (RFC-invariance in Brown, 2007). Rescaling all $x \rightarrow \alpha x$, with α positive, results in $W \rightarrow \alpha W$. This means that the ordering between histories h and g remains the same after rescaling. This invariance is important when the unit of well-being is not fixed. It is like comparing the lengths of two sticks in terms of meters or centimeters.

When the weights a_i are uniform (equal to 1), the welfare function has a third important property: permutation symmetry of the well-being levels of the different individuals. Reordering the individuals gives the same welfare function. This means that all individuals are treated impartially. When the weights are not uniform, we can reformulate the theory to keep it impartial. Write the lifetime well-being vector as $x_\uparrow = (x_{[1]}, x_{[2]}, \dots)$, i.e. in ascending order ($x_{[1]} \leq x_{[2]} \leq \dots \leq x_{[N]}$). In this interpretation, $x_{[l]}$ does not refer to the well-being of individual l , but to the l -th level of well-being.

4.6.4 Deriving the welfare function behind the veil of ignorance

The parameters in the above proposed welfare function will be determined behind the veil of ignorance, by borrowing some elements from prospect theory (Kahneman & Tversky, 1979).

A first important element in prospect theory is the reference point relative to which losses and gains are measured. The lifetime well-being is the value you as impartial observer would ascribe to the complete life of a sentient being, i.e. your preference for living that life. A lifetime well-being higher (lower) than 0 is a life that is (not) worth living. You prefer not being born to being born as someone with a negative lifetime-well-being. So we can assume that living a life of value 0 is equally preferable to not

being born at all²¹, to living a completely unconscious life without any experiences, or to living a conscious life without any positive and negative feelings or preferences (these three options are equal in terms of well-being, although only the latter, conscious life is the life of a sentient being and is included in the welfare function). The zero value is the reference point, and positive (negative) lifetime well-being levels are considered as gains (losses) behind the veil.

As the impartial observer does not know the identity of the individual that s/he will be in the real world, s/he does not know what lifetime well-being s/he will get. For the impartial observer, it becomes a game of chance. In the next three sections, I apply three elements from prospect theory, to our game of chance: risk aversion towards gains, the reflection effect (no risk aversion towards losses), and loss aversion. Next, I also discuss uncertainty aversion.

4.6.4.1 Risk aversion for positive well-being levels

Suppose the probability distribution is uniform: you have an equal probability of being born as any sentient being. If there are N sentient beings in time interval Δt , then your probability of being individual i is $1/N$. Due to impartiality, this uniform probability distribution implies uniform weights a_i in the welfare function.

When the possible outcomes are gains, people often tend to be risk averse (Kahneman & Tversky, 1979). Suppose there are two histories, both having two individuals. In history h_1 the individuals have well-being $x_1(h_1) < x_2(h_1)$. Behind the veil of ignorance, you don't know whether you will get the lower or the higher level of

²¹ This assumes that a non-existing life has a zero lifetime well-being. It requires that lifetime well-being is an 'extensive' instead of an 'intensive' quantity. In physics, examples of extensive quantities are lengths, masses, energies and electric charges, whereas densities, pressures, temperatures and chemical potentials are examples of intensive quantities. Also averages are intensive quantities. Consider the example of me holding a non-existing cup of tea in my hand. The question how much tea I hold in my hand is well defined, even if the cup does not exist: 0 liter. In contrast, the question what the average amount of tea is per cup in my hand, is not well defined: 0 liter divided by 0 cups is mathematically not defined. Such an average is an intensive quantity, and intensive quantities cannot properly deal with non-existence. I believe that lifetime well-being is an extensive quantity, which means that it is defined in cases of non-existence. But the welfare function is an intensive quantity, as it is an average over sentient beings. This implies that we can compare a life that has a well-being with the absence of that life. But we cannot compare the value of a world that contains sentient beings with a world without sentient beings. As the welfare function is an average, a non-sentient or non-existing being adds 0 lifetime well-being to the numerator and 0 to the population size in the denominator. On the other hand, an existing, sentient being that has 0 lifetime well-being adds 0 to the numerator but 1 to the denominator. Hence, in the denominator we can see the difference between non-existing, non-sentient beings and existing, sentient beings.

well-being, so it becomes a game of chance. In history h_2 there is no gambling, which means that $x_1(h_2) = x_2(h_2)$.

In the theory of risk aversion (Arrow, 1965), there is the notion of a utility function $u(x)$ which allows comparison of both histories. This utility function can be derived by using the condition that when the two histories are equivalent, the average utilities in both situations should be the same. Hence, if the welfare function represents your preference for a certain history, then we first have to ask what is the relation between the levels of well-being in both histories such that you would have an equal preference for both histories. In other words: when would the certain outcome in history 2 be equal to the game of chance of history 1? Equality of the welfare function $W(x(h_1)) = W(x(h_2))$ solves this question. This gives

$$x_1(h_2) = x_2(h_2) = \left(\frac{x_1(h_1)^p + x_2(h_1)^p}{2} \right)^{\frac{1}{p}}$$

Using the condition that when the two histories are equivalent, the average utilities in both situations should be the same, we get as a condition:

$$\frac{u(x_1(h_1)) + u(x_2(h_1))}{2} = \frac{u(x_1(h_2)) + u(x_2(h_2))}{2} = u\left(\left(\frac{x_1(h_1)^p + x_2(h_1)^p}{2}\right)^{\frac{1}{p}}\right).$$

We see that this equality is solved when the utility function is identical to the power function: $u(x) = x^p$.

The Arrow-Pratt measure of relative risk aversion (Arrow, 1965; Pratt, 1964) with respect to the well-being is defined as:

$$R(x) \equiv -x \frac{u''(x)}{u'(x)} = 1 - p$$

This relative risk aversion is constant, which means that the utility function is so-called iso-elastic. When $p=1$, the relative risk aversion is 0 and we get average utilitarianism. When $p \rightarrow -\infty$, the risk aversion goes to infinity, and we get a maximin welfare function. In between these two extremes, a non-zero and non-infinite risk aversion behind a veil of ignorance results in prioritarianism.²²

²² Note that when $p \leq 0$, then $\langle x^p \rangle_p = 0$ as soon as there is at least one x_i less or equal to 0. As a result, once there is at least one being with zero or negative lifetime well-being, the positive part of the welfare function either becomes trivially zero, or the risk aversion parameter p should be higher than 0. The latter restriction on the risk aversion implies that we cannot move close to a maximin theory. One could try to take a power p that depends on the levels of well-being (i.e. $p=p(x)$), such that $p>0$ when there are individuals having a well-being at or below 0. Another, perhaps more elegant option to avoid this problem is to take an exponential function of the lifetime well-being, instead of a power root function (see Lumer, 2006). Take the exponential lifetime well-being

4.6.4.2 Priority for the worst-off

For positive well-being levels, the first term of the welfare function corresponds with the priority view (Scheffler, 1982; Weirich, 1983; Parfit, 1991; Parfit, 1997; Holtug, 2006): in distributing benefits to individuals and maximizing lifetime well-being, the worst-off individuals should get a priority. Mathematically, this can be expressed as the Pigou-Dalton principle: transferring a quantity Δx from the better-off (lifetime well-being x_1) to the worse-off (lifetime well-being $x_2 < x_1$), without reversing the order (i.e. $x_2 + \Delta x < x_1 - \Delta x$) increases the welfare function.

Note that the welfare function for positive well-being levels has maximin and sum-utilitarianism as limits. When the power $p \rightarrow 1$, the welfare function becomes a sum of well-being levels, which means that no-one has a priority, as in sum-utilitarianism. When the power p goes to minus infinity, the theory becomes maximin, where the worst-off individual gets absolute priority for its well-being. The latter is not true for traditional expressions of concave prioritarianism, where the welfare function $W = \sum f(x)$ with f a concave function (as in Broome, 1991; Brown, 2007; Holtug, 2006; Lumer, 2006; Rabinowicz, 2002; McCarthy, 2008). The power $1/p$ in our welfare function allows us to take this limit $p \rightarrow -\infty$. When p is very negative, we get a quasi-maximin prioritarianism.

In the previous section (4.4), I mentioned a second justification for prioritarianism, next to impartiality (the veil of ignorance) with a high but not maximal risk aversion: empathy with a low, but not zero need for efficiency. The efficiency can be measured by looking at extended Pigou-Dalton transfers (Vallentyne, 2009, p.158). Such a transfer is given by the inequality conditions: $x_1(h_1) \leq x_1(h_2) \leq x_2(h_2) \leq x_2(h_1)$. This means

$$f_a(x) = \frac{a}{a-1} (1 - a^{-x})$$

instead of $f_p(x) = x^p$, with the parameter $a \geq 1$. If this parameter a equals 1, we get (number-dampened) sum-utilitarianism because $f_1(x) = x$. In the limit of the parameter going to infinity, the welfare function becomes maximin. The positive part of the welfare function now becomes a (number-dampened) Kolmogorov mean (Kolmogorov, 1930) with exponential functions:

$$W_{f_a+}(x(h)) = \frac{N_+}{N_+ + N_R} f_a^{-1} \left(\frac{1}{N_F} \sum_{i=1}^{N_F} f_a(x_i^+) \right),$$

with

$$f_a^{-1}(y) = - \frac{\ln \left(\left(\frac{1-a}{a} \right) y + 1 \right)}{\ln a}$$

the inverse of the exponential function. This new welfare function has a priority for the worst-off and reflects a constant absolute risk aversion, defined as

$$A(a) \equiv - \frac{f''(x)}{f'(x)} = \ln(a).$$

that switching from history h_1 to h_2 , the well-being of the lowest level increases and the well-being of the highest level decreases (but the order of the levels doesn't change). The extended Pigou-Dalton transfer efficiency E is determined by the ratio of the benefit for the lowest level to the cost for the highest level, where benefits and costs mean increases and decreases in well-being:

$$E \equiv \frac{x_1(h_2) - x_1(h_1)}{x_2(h_1) - x_2(h_2)}.$$

If we take small, neutral extended Pigou-Dalton transfers, i.e. infinitesimal transfers with $W(h_1)=W(h_2)$, then using the derivatives of the power function evaluated in $x_1(h_2)$ and $x_2(h_2)$, the efficiency becomes approximately

$$E \approx \left(\frac{x_1(h_2)}{x_2(h_2)} \right)^{1-p}.$$

When $p=1$, the efficiency is always 1, i.e. maximal. When relative risk aversion $1-p$ increases, the efficiency decreases. Infinite risk aversion always corresponds with zero efficiency, which results in maximin prioritarianism.

4.6.4.3 Avoiding the replaceability problem

An objection to a total utilitarian theory (taking a sum of the well-being of all individuals) is that sentient beings are treated as nothing but receptacles of well-being. In total utilitarianism it is not a moral problem if a person is simply replaced by another person with the same level of well-being. For example, you are allowed to kill someone as long as you let another person be born, who will have the same expected well-being as the murdered person would have if s/he was not killed. This is counter-intuitive.

One can counter this replaceability problem by simply stating that persons have a unique intrinsic value, which simply means that these persons cannot be replaced without violating something deemed important. However, an advantage of using the lifetime perspective is that we can avoid this replaceability problem without a need to introduce such an intrinsic irreplaceability value. Due to the lifetime perspective, the number N_F of individuals over time is well defined. Hence, we can construct a welfare function that uses this number of individuals. We can use for example an average instead of a total of (priority weighted) well-being (i.e. we can include a division by N_F) or include a number-dampening factor in the welfare function.

As an example, compare situation $h=(100)$, i.e. a situation where one person has a lifetime well-being equal to 100, with situation $g=(50;50)$, i.e. a situation where one person is killed somewhere in the middle of his life (so he has lifetime well-being 50), and is replaced by a second person who will get the remaining lifetime well-being of 50. According to total utilitarianism, both situations are equally good. But applying the

prioritarian theory, we see that $W(g) < W(h)$. Hence, the person should not be killed and replaced. Due to the lifetime perspective we can avoid the replaceability problem.²³

4.6.4.4 Avoiding the non-identity problem

The non-identity problem (Parfit, 1984) asks questions like: Can we harm someone if the other option we had would be that this sentient being would not have existed? Or can we harm a future being by bringing that being into existence (e.g. breeding an animal in the livestock industry)? The problem is that both yes and no answers somehow violate our moral intuitions. It is difficult to imagine how we can harm someone if we'd say to that person: "But if you didn't want to be harmed, it means you'd prefer not to exist." On the other hand, it also seems wrong to let someone be born, knowing that this person will suffer tremendously.

The reason why the non-identity problem is avoided in our QMM-prioritarian theory is because the QMM-theory is actually not about harming someone, but it's about just distributions of lifetime well-being. With the above welfare function, we can simply avoid the tricky questions raised by Parfit, because I didn't say anything about the identity of the different persons. It doesn't matter if person 1 in situation X equals person 1 in situation Y.²⁴

²³ Average utilitarianism combined with the lifetime perspective also avoids the replaceability problem, because in average utilitarianism the welfare function is divided by the number of individuals, and we get $100/1 > (50+50)/2$.

²⁴ The QMM-theory uses a notion of 'impersonal harm' (Parfit, 1984, p.387) or 'wide person affecting harm' (Visak, 2011). One could argue that every notion of harm requires an impersonal or wide person affecting view. Consider diachronic harm: A harms B at time t if B gets a lower momentaneous well-being after t than before t . But as discussed in section 4.2, the notion of a personal identity over time is tricky and vague. In fact, B might become a (slightly) different person after time t , so A harms a different person. The same applies to a counterfactual notion of harm: A harms B at time t if the well-being of B after t in a world where A does act is lower than the well-being that B would have got after t in a parallel, counterfactual world where A did not act. As with the Parfitian thought experiments of copying and splitting persons, we can say that person B in the real world is a (slightly) different person than person B' who lives in the counterfactual world. An example: A builds a house for B. The house is quickly built, which means that B is able to move in the house a month earlier. As a consequence, that first month in his new town, B happens to meet a girl who will later become his wife and who will change his life (and personality) profoundly. But as the house was quickly built, it has a weaker construction, and at one day collapses and kills B. In the parallel world, B' would have to wait another month before he can move in the house. As a consequence, B' does not meet the girl and he will live a completely different future life. But the house does not collapse and B' is not harmed. The question is: at the moment of collapse, is B in the first world the same person as B' in the counterfactual world? If not, then the non-identity problem already occurs in counterfactual notions of harm, and the narrow person affecting view (Visak, 2011) would run into counterintuitive troubles.

4.6.5 The reflection effect and risk neutrality for negative well-being levels

According to the reflection effect of prospect theory (Kahneman & Tversky, 1979), the risk attitude towards losses is different than the risk attitude towards gains. For negative well-being levels, people behind the veil of ignorance are no longer risk averse; they become more risk seeking. As the parameter p determined the risk attitude for positive well-being levels, the welfare function for negative well-being levels is dependent on a similar parameter q . When $q < 1$, we get risk seeking behavior for losses. This corresponds with the proposal of triage as discussed in Brown (2007). However, triage can be counter-intuitive, as it gives priority to the better-off of the negative well-being levels. Therefore, and for simplicity, I take risk neutrality for losses. This corresponds with $q = 1$, i.e. total utilitarianism for negative well-being levels.

4.6.6 Loss aversion

According to prospect theory, people have a tendency to prefer avoiding losses to acquiring gains (Kahneman & Tversky, 1984). Loss aversion can be introduced in the welfare function, by taking the parameters $s = 1$ and $t = 0$ in the number-damping functions.

Consider a history h with a population divided in two equal subpopulations $N_+ = N_-$ and for each individual i in the positive population there is a corresponding individual j in the negative population with: $x_j^- = -x_i^+$. In other words: the well-being levels are distributed symmetrically around the reference value 0.

Looking at the welfare function for this situation, noting that $\langle x^+ \rangle_p \leq \langle |x^-| \rangle_1$ and $P_{1,N_R}(N_+) < P_{0,N_R}(N_-)$, we get $\mathcal{W}(h) < 0$. This indicates a loss aversion: one would rather have a well-being 0 with certainty, than taking a gamble with a symmetric distribution of well-being levels around zero, because the possible losses count heavier than the gains.

This loss aversion is consistent with some moral intuitions in population ethics. Due to loss aversion, and in particular the choice of the parameters $s = 1$, we can avoid the repugnant and reverse repugnant conclusions. The choice of $t = 0$ allows us to avoid the strong sadistic conclusion. This will be explained in the next sections.

4.6.6.1 Avoiding the repugnant conclusion

The Repugnant Conclusion (Parfit, 1984) is one of the most challenging arguments in population ethics. The argument goes as follows. Start with a population of very happy people who have well-being 100. So, situation $A = (100)$. The total-utilitarianist would say

that a situation $B=(100;98)$ is better than A : the total number of happy persons increased, and the total happiness increased by adding people who are only slightly less happy than the existing population. This addition of happy people is the first step. The second step consists of equalizing the levels of well-being: situation $C=(99;99)$ is considered to be even better than B , because now there is more equality. One can repeat step 1, by introducing a third population with well-being levels equal to 97. After repeating step 1 and step 2, we move to the optimal situation Z which contains an almost infinite number of people with an almost zero (but still positive) well-being. Each of those individuals still has a life that is worth living, because their lifetime well-being remains positive, though very low. So Z is better than A , or in other words, we should boost population growth, even if all well-being levels become very low. But this conclusion is repugnant.

The first term in the proposed welfare function shows how the repugnant conclusion can be avoided: for large populations N , the population factor $P_{1,N_R}(N_+)$ becomes constant and we end up with the power mean of well-being levels. Adding more and more persons with lower and lower well-being, lowers this power mean, and hence lowers the welfare function.

As a side remark, the repugnant conclusion might also be avoided in total utilitarianism, i.e. when $p=1$, when three conditions are met.²⁵ First, an individual lifetime well-being is dependent on a resource with decreasing marginal utility. Hence, the lifetime well-being can be written as a concave function $f(r)$ of a resource r available to the individual. Second, there is a critical consumption level c of the resource required for a positive well-being. And third, the total amount of the resource R is limited. If all individuals have an equal access $r=R/N$ to the resource, then we can write well-being for each individual as

$$x_i^+ = \sqrt{\frac{R}{N} - c}.$$

The welfare function $W = Nx_i^+$ is maximal for a finite, optimal population $N_+^{opt} = R/2c$. In this optimum, the lifetime well-being of an individual becomes \sqrt{c} . Under these realistic conditions, the population size does not explode as in the repugnant conclusion. Nevertheless, a worry remains: what if the average well-being \sqrt{c} is very

²⁵ Shiell (2005) gave a more general proof that total utilitarianism avoids the repugnant conclusion when there are five restrictions which reflect universal properties of physics, biology and preferences: essentiality of material consumption (the fact that x depends on r), positive subsistence consumption (a positive critical consumption level c), upper bounds on resources (R) and non-material goods, and the law of conservation of matter.

low? This optimum might be repugnant enough in total utilitarianism. To play it safe, I prefer the power mean prioritarianism to avoid the repugnant conclusion.

4.6.6.2 Avoiding the reverse repugnant conclusion

If using a power mean of well-being levels solves the repugnant conclusion, we have to be aware of a reverse repugnant conclusion. Maximizing a power mean implies that it is not good to give birth to beings who will get a lower well-being than the power mean of well-being of the existing individuals. So if we systematically exclude births of potential beings who will have lower well-being levels, in the end, only the person with the highest well-being should be born. Instead of overpopulating the world, an average well-being welfare function says that we should underpopulate the world.

Due to the choice for $s=1$, the population factor in the positive term of the welfare function is a concave function of the population size N_* .²⁶ That means that for low populations (lower than the reference value N_R) the welfare function increases linearly in N_* . This pulls us away from an under-populated world, because it is good to increase the population size.²⁷

4.6.6.3 Trade-off between quantity, quality and equality

As the priority view lies between maximin and utilitarianism, it represents a trade-off between equality and efficiency. According to maximin, inequality is only allowed if it benefits the worst-off. According to utilitarianism inequality is always allowed as long as the distribution of well-being is efficient, i.e. as long as total well-being is maximized.

²⁶ Note that this adapted moral weight expression has a similar structure as the value of life expressed in section 4.2.4, generating a trade-off between quantity (number of individuals, or length of a lifetime of a single individual) and quality (values of life of different individuals, or experienced well-being of a single individual).

²⁷ Another way to avoid the reverse repugnant conclusion, is by introducing a deontological permission that is related to the 3-N-principle to be discussed in section 1.1(10.4). Suppose the welfare function simply contains a generalized f-mean (Kolmogorov, 1930)

$$W_{f_a}(x(h)) = f_a^{-1} \left(\frac{1}{N_F} \sum_{i=1}^{N_F} f_a(x_i) \right)$$

with the weight function for example the exponential lifetime well-being (Lumer, 2006):

$$f_a(x) = \frac{a}{a-1} (1 - a^{-x}).$$

This welfare function does not include a population factor. It is simply an average, so the repugnant conclusion is avoided, but the reverse repugnant conclusion is not avoided. However, we can add a deontological permission which says that everyone is allowed to procreate, even if the new individuals would get a lifetime well-being lower than the generalized f-mean of all other individuals. As we will see in section 10.6, procreation is always allowed, because procreation is natural, normal and necessary for e.g. biodiversity.

Adding the population factor $P_{1,N_R}(N_+)$, we get a trade-off between three elements: quantity (population size), quality (efficiency) and equality. These three elements correspond with three factors in the welfare function²⁸:

$$W_+(x(h)) = \frac{N_+}{N_+ + N_R} \cdot A(x^+) \cdot (1 - I_A(x^+)),$$

with $A(x^+) = \langle x^+ \rangle_1$ the average well-being and

$$I_A(x^+) = 1 - \frac{\langle x^+ \rangle_p}{\langle x^+ \rangle_1} \in [0,1]$$

the Atkinson inequality index (Atkinson, 1970; Sen, 1982). When $p < 1$, this inequality index is zero only when all x_i are equal.

4.6.6.4 Avoiding the strong sadistic conclusion

The negative mere addition principle (Arrhenius, 2000, p. 66)²⁹ says that adding individuals with a negative lifetime well-being always lowers the welfare function. However, consider a negative average utilitarianism, with average well-being -100. In such a theory, adding someone with lifetime well-being -1 would increase this average, and hence would be an improvement. This is the strong sadistic conclusion of negative average utilitarianism: adding someone whose life is not worth living might increase the welfare function.

Our welfare function contained a negative term $-P_{t,N_R}(N_-)\langle |x^-| \rangle_q$. If the parameter $t=0$, then the strong sadistic conclusion is avoided, because in this case the negative term does not have a N_- in the denominator. If the negative term would have a N_- in the denominator of the population factor, the term can increase (become less negative) when N_- increases, i.e. when people are added whose lives are not worth living.³⁰

²⁸ This expression of the welfare function gives 2-dimensional indifference surfaces in a 3-dimensional space, as discussed in Carter (1999).

²⁹ This corresponds with the “Hell Three” thought experiment in Parfit (1984, p. 422).

³⁰ Another way to avoid the strong sadistic conclusion, is by introducing a deontological constraint that is related to the mere means principle to be discussed in section 6.2. Suppose the welfare function simply contains a generalized f-mean (Kolmogorov, 1930)

$$W_{f_a}(x(h)) = f_a^{-1} \left(\frac{1}{N_F} \sum_{i=1}^{N_F} f_a(x_i) \right).$$

This welfare function includes the negative levels of well-being and does not include a population factor. Hence it is vulnerable to the strong sadistic conclusion when the generalized f-mean is negative: the welfare function increases when a new individual is added who has a negative lifetime well-being which is higher than

4.6.6.5 Priority for negative levels of well-being

The welfare function is discontinuous around 0: when a positive well-being level decreases till it reaches the zero value, the population factor suddenly drops from $P_{1,N_R}(N_+)$ to $P_{1,N_R}(N_+ - 1)$, because the number positive well-being levels decreases. Similarly, the population factor $P_{0,N_R}(N_-)$ suddenly increases if the well-being further drops below zero.

The discontinuity of the population factors might seem counter-intuitive to some people, but a nice feature is that the welfare function now contains a sufficientarian (critical threshold) element: lifting a negative well-being up to a positive well-being becomes very important.

Note that avoiding the discontinuity by taking a population factor $P_{1,N_R}(N_F)$ instead of $P_{1,N_R}(N_+)$ generates stronger counter-intuitive problems, such as a strong sadistic conclusion. If the population factor contains the total number of beings N_F , and if N_F is much lower than N_R , then the welfare function might increase when a new being with a small negative well-being is introduced.

Note also that the power mean $\langle x^+ \rangle_p$ contains the total number of beings N_F instead of only N_+ . If the power mean was restricted to the positive well-being levels, another very strong sadistic conclusion might occur: decreasing someone's well-being till it reaches the zero value might suddenly increase the welfare function.

4.6.6.6 Preference for (not) being born

The choice of the population factors $P_{1,N_R}(N_+)$ and $P_{0,N_R}(N_-)$ generates loss aversion relative to the zero well-being reference. These factors can also have a very different interpretation, not related to loss aversion. We can formulate the veil of ignorance in such a way that the population factors correspond with conditional probabilities of being born as an individual with a positive or negative lifetime well-being. This conditional probability is related to the question: how many potential beings are sitting behind the veil, and how many of them will be born in front of the veil?

Suppose you are an impartial observer behind the veil of ignorance. The number N_F of beings actually born in *front* of the veil in time interval Δt , depends on your choice of history h . For the number of beings behind the veil, there are three options.

this generalized f-mean. But this new individual will be used as merely a means to increase the welfare function: the presence of the individual is required and the individual does not want to live because s/he has a negative lifetime well-being.

Option 1: you are as impartial observer alone behind the veil, and after your choice of world history, $N_F - 1$ extra beings are created, so $N_B = N_F$. In this case, you are certain to be born.

Option 2: the number N_B of potential beings behind the veil is already determined before you, as impartial observer, choose a certain history. N_B can be written as $Z\Delta t$, with Z the maximum possible birth rate. Of those N_B potential beings, only N_F will be born. Your probability to be born might be very low, because N_B might be very high.

Option 3: something in between the previous two option: there is an infinite pool of all possible beings. Once a world history h is chosen, a number $N_B \geq N_F$ is drawn from the pool. N_F of them will actually be born. Your probability to be born is now N_F/N_B , which might be low, but not as low as in option 2.

Suppose first that all well-being levels are positive or zero. According to the first formulation, you are sure to be born, so you can as well maximize the (power) average well-being:

$$W_1 = \langle x^+ \rangle_p.$$

This expression faces the reverse repugnant conclusion.

According to the second formulation, you can increase your probability of being born by increasing the actual population size N_F . The welfare function reads

$$W_2 = \frac{N_F \langle x^+ \rangle_p}{Z\Delta t},$$

with the denominator $Z\Delta t$ a constant (independent from N_F). This expression faces the repugnant conclusion.

The third option avoids both repugnant conclusions. Not knowing whether you have to maximize total or average weighted well-being, we could take a combination, such as:

$$W_3 = \frac{N_F \langle x^+ \rangle_p}{N_B},$$

with $N_B = N_F + N_R$. This equals number-dampened prioritarianism.

When some well-being levels are negative, the third option can be reinterpreted. You first get a probability $N_+/(N_+ + N_R)$ of being born with a positive well-being. You want to maximize this probability, but there is a trade-off with the quality, i.e. the power average well-being. If you are not selected to have a positive well-being, you are left with a probability of being born with a negative well-being equal to N_-/N_R . This equals the population factor for negative well-being levels. Note that this is a conditional probability (conditional on not being born with a positive well-being), so it no longer contains a term N_+ in the denominator. The number of unborn beings equals $N_U = N_R - N_- - N_0$. This number is variable, such that N_R can be treated as a constant (independent from N), to avoid the strong sadistic conclusion. Note that if N is big, N_R should also be big. That means that when there are a lot of individuals with negative well-being levels, the population factor for positive well-being levels $P_{1,N_R}(N_+)$ increases almost linearly

in N_+ . In other words: the more individuals whose lives are not worth living are born, the more individuals should be born whose lives are worth living.

We see that the population factors $P_{1,N_R}(N_+)$ and $P_{0,N_R}(N_-)$ have three different justifications: 1) they correspond with moral intuitions to avoid the repugnant, reverse repugnant and strong sadistic conclusions, 2) they correspond with a loss aversion behind a veil of ignorance and 3) they correspond with conditional probabilities in a differently constructed veil of ignorance thought experiment. These different perspectives generate a coherent picture.

4.6.6.7 Summary: positive number-dampened power mean prioritarianism and negative total utilitarianism

For uniform probability distributions ($a_i = b_j = 1$), the reflection effect combined with risk and loss aversion allows taking the parameters $p \in]-\infty, 1[$, $q=1$, $s=1$ and $t=0$. Then the welfare function reads:

$$W(x(h)) = \frac{N_+}{N_+ + N_R} \left(\frac{1}{N_F} \sum_{i=1}^{N_F} (x_i^+)^p \right)^{\frac{1}{p}} + \frac{N_-}{N_R} \left(\frac{1}{N_F} \sum_{j=1}^{N_F} x_j^- \right).$$

This expression has two parameters (p and N_R) that can be tuned to correspond with our intuitions.

The first term of the welfare function is a *number-dampened power mean prioritarianism*. Here, prioritarianism refers to a welfare function with a sum of concave utility functions such as $u(x)=x^p$, with $p \in]0,1[$ (Broome, 1991; Brown, 2007; Holtug, 2006; Lumer, 2006; Rabinowicz, 2002; McCarthy, 2008). The power mean prioritarianism is a generalization to negative powers (this generalization is made possible due to the overall reverse power $1/p$ in the welfare function). The number-dampened property refers to the extra factor that is linear in N_+ for very small populations and nearly constant for very large populations (see number-dampened utilitarianism: Hurka, 1983; Blackorby et al. 2002; Ng, 1986).

The second term of the welfare function corresponds with a kind of *negative total utilitarianism*, a theory where only negative lifetime well-being levels count, and where the only objective is to decrease this total amount of negativity.

This welfare function corresponds with some moral intuitions in population ethics. It gives a priority for the worst-off (and a special priority for the beings with a negative well-being), generates a trade-off between quantity, quality and equality, and avoids the replaceability problem, the repugnant conclusion, the reverse repugnant conclusion and the strong sadistic conclusion.

However, in population ethics, a lot of impossibility theorems are proven (see e.g. Arrhenius, 2000; Blackorby et al. 2003; Brown, 2007). That means that our welfare

function cannot escape some problems. In fact, it faces three troublesome, counter-intuitive conclusions.

4.6.7 Problematic properties of number-dampened prioritarianism

4.6.7.1 Independence and the mere addition paradox

Roughly speaking, independence or strong separability (see e.g. McCarthy, 2008) says that the choice that maximizes the welfare function should not depend on individuals whose presence and well-being levels cannot be influenced. This means that the moral judgment of a change that only affects a subpopulation does not depend on the rest of the population. More accurately: suppose we have to decide between two histories h and g , and in h there is an unaffected subpopulation of N_{un} people who have the same levels of lifetime well-being as the corresponding N_{un} people in history g . This subpopulation consists of the unaffected people, because in both choices their well-being is the same. Imagine this subpopulation on a far away island, outside the influence of our choices. Now we transform h and g to h' and g' respectively. For people *not* in the subpopulation, i.e. for people in the affected population of size $N_{af} = N_{tot} - N_{un}$, individuals in h' have the same well-being as individuals in h , and the same goes for g' and g . But for the people in the unaffected subpopulation, the well-being levels are changed in the same way for both histories: for the subpopulation, the transformation from h to h' is the same as the transformation from g to g' . So, for example the well-being of all people in the unaffected subpopulation is raised (e.g. the people on the far away island discovered new resources). In that case, independence says that $W(h) > W(g)$ if and only if $W(h') > W(g')$, i.e. the order of preference should not change.

The principle of independence (strong separability) is valid in our prioritarian theory in situations where the numbers of people with positive and negative levels of well-being remain constant. However, the principle is violated in two cases: in mixed populations (where choices can influence the numbers of people with positive and negative well-being) and in variable populations (where choices can influence the total number of people).

Consider mixed populations first: suppose that in history h the affected subpopulation has some people with positive and some with negative lifetime well-being levels, whereas in g the affected population has different numbers of people with positive and negative levels. For example some people with positive well-being in h get a negative well-being in g . Suppose that the unaffected people in h have positive levels of well-being, whereas they have negative levels in h' . In that case, it is easy to

demonstrate that $\mathcal{W}(h) > \mathcal{W}(g)$ does not necessarily result in $\mathcal{W}(h') > \mathcal{W}(g')$, because the unaffected population influences the population factors.³¹

If most of our policy choices do not flip the sign of someone's well-being (i.e. do not change a positive into a negative well-being or vice versa), this violation of independence is not a serious problem. The violation only implies that the amount of priority for negative levels of well-being (how important it is to raise someone's negative well-being instead of raising someone else's positive well-being) can depend on unaffected populations.

As most policy choices can influence the population size, there is a second, perhaps more serious kind of violation of independence in situations with variable populations. Take for example the power $p=1/2$, N_R very low and consider two times two histories, each with populations having well-being levels $h_1=(1)$, $h_2=(1;9)$, $h'_1=(1;81)$ and $h'_2=(1;9;81)$. We see that history h'_1 is related to history h_1 , just as history h'_2 is related to h_2 : I simply added a population (having a well-being 81 per individual) to both histories h_1 and h_2 in order to get histories h'_1 and h'_2 . We can imagine that this happy population is added on a far away island, so we expect that this population does not influence our choices here. The choice we face is between a situation with a population having well-being 1, versus a situation with a doubled population, whose individuals have well-being levels 1 and 9. However, $\mathcal{W}(h_1) < \mathcal{W}(h_2)$ but $\mathcal{W}(h'_1) > \mathcal{W}(h'_2)$. In other words: the existence of a population on a far away island might reverse the order of preference here.

These violations of independence are counter-intuitive. Nevertheless, restoring independence would result in the violations of two other moral intuitions: the preferences that an impartial observer has behind the veil of ignorance and the intuition that we should avoid the repugnant conclusion. The latter two moral intuitions might be stronger (have more coherence) than the single moral intuition of independence. If that is the case, then constructing the most coherent reflective equilibrium would imply that it is better to violate independence.

Related to this violation of independence for variable populations is the mere addition paradox. We encountered this paradox in the first step towards the repugnant conclusion: how can it be bad to add very happy people? In order to avoid the repugnant conclusion, this badness came down to the observation that the average well-being decreases when people are added who are (slightly) less happy than the

³¹ Note that a generalized f-mean (Kolmogorov, 1930)

$$W_{f_a}(x(h)) = f_a^{-1} \left(\frac{1}{N_F} \sum_{i=1}^{N_F} f_a(x_i) \right).$$

that includes the negative levels of well-being and does not include a population factor, has independence for mixed populations (but not for variable populations).

existing group. By moving from $A=(100)$ to $B=(100;98)$, no-one is harmed, the total well-being almost doubles, and all lives are worth living. For small populations, such additions are good according to our welfare function. Yet, such additions are not good when N_+ becomes larger than the reference N_R . For large populations, quality dominates over quantity.

It seems paradoxical that merely adding happy people could lower the welfare function. The paradox gets worse when we imagine far away planets with trillions of super happy extraterrestrial beings. Our human happiness cannot compete with theirs, so we only lower the average well-being if we keep on procreating. It would be better if humans stopped procreating. The same goes for a world history time interval Δt that includes a past era of dinosaurs: what if scientists discovered that they were happier than humans? How is it possible that the ethics of human procreation would depend on such discoveries?

Three remarks are in order. Combining those three remarks will sufficiently weaken the problem of the mere addition paradox.

First, a weak remark: this paradox only occurs in situations when populations are already large. The mere addition paradox is avoided in small populations, due to the population factor $P(N)$ that becomes linear in N . However, for large populations (or if populations include all far away planets), this is not much of a consolation.

Second, and more importantly, we can soften the contra-intuitive mere addition paradox by writing the welfare function as a sum of a changeable and an additional part: $W = W_{ch} + W_{add}$. The changeable part looks like the expression presented above and contains individuals whose lifetime well-being can be changed amongst each other. The additional part is simply the sum of lifetime well-being of additional beings. The point is that this additional part contains only those beings whose lifetime well-being cannot be exchanged with beings from the changeable group. In particular it is impossible to transfer well-being from the changeable to the additional group. Adding individuals who have a lower lifetime well-being than the critical level determined by W_{ch} (e.g. lower than the population weighted, power mean of the levels of well-being of the changeable group) will increase the total welfare function if those individuals belong to the additional group. Writing the welfare function like this avoids the repugnant conclusion, because the argument towards the repugnant conclusion breaks down at the second step: a move from $A=(100)$ to $B=(100;98)$ is allowed (satisfying mere addition), but the second move from $B=(100;98)$ to $C=(99;99)$ is impossible because the second person belongs to the additional group.

Third, and most importantly, even though adding people who belong to the changeable group and who have a too low lifetime well-being is not good according to the welfare function W_{ch} , we can include some deontic permissions. These are permissions that are always allowed, even when they violate the prioritarian welfare ethic.

Most people have moral intuitions about three such deontic permissions. They become particularly visible when we include non-human animals as sentient beings in the welfare ethic. The first deontic permission says that predation is allowed (we do not have a duty to protect prey from predators), even when predators violate the welfare ethic by killing a lot of prey. Second, animals are allowed to move around, even if small insects would have a well-being that is lowered when they are in huge numbers trampled by the large animals. And third, procreation is allowed, even if an animal species does not contribute enough to the welfare function. Compare a frog with a low lifetime well-being (a poor emotional life over a short, eight year lifespan, with a low psychological connectivity) with a normal human with a high lifetime well-being (a rich emotional life over a longer, eighty years lifespan). The frog still has a life worth living: his lifetime well-being is positive.

Two explanations might justify such deontic permissions. First, biodiversity might have a moral value. If frogs (and all other life forms with lower lifetime well-being) were not allowed to procreate when happier humans exist (or if humans are not allowed to procreate when happier ET's or dinosaurs existed), then biodiversity would drastically decrease. This decrease in biodiversity trumps the prioritarian welfare ethic. The same goes for the extinction of all predators when they are not allowed to prey on animals and for large animals when they are not allowed to move and kill insects by accident. A second solution is a kind of behavioral fairness: if humans are allowed to procreate, then so are frogs. If zebras are allowed to eat for survival, then so are lions. If insects are allowed to move, then so are elephants.

Both arguments of biodiversity and fairness can be combined into the following deontic permission of procreation: if person X is allowed to do something that necessarily contributes to biodiversity (such as procreation), then so is person Y, even if the welfare function decreases (i.e. even if Y's child has a positive lifetime well-being below the power mean). These explanations of deontic permissions can be further refined, but I will leave that to a later chapter on the predation problem (Chapter 10). Here, it is sufficient to note that there might exist such deontic permissions that are coherent with moral intuitions about e.g. predation, motion and procreation.

In summary, even if some behavior such as procreation would lower the welfare function, it is always allowed (but not obligatory)³². Welfare functions are applicable to different, incomparable regimes, corresponding to different populations. Once new beings with lower well-being are added, we enter a new regime. Such a shift from regime is always permissible. In this new regime, we have to move on maximizing the

³² Some restrictions might be included: perhaps it is never permitted to add an individual with a negative lifetime well-being. Hence, species who can only have negative well-being levels are not allowed to procreate.

welfare function. Killing the added beings would not be an improvement: it is impossible to return to a previous regime.

4.6.7.2 The weak sadistic conclusion

Arrhenius (2000, p65) pointed out that a theory such as number-dampened prioritarianism is vulnerable to a weak version of the sadistic conclusion. As mentioned, the addition of people with low positive well-being levels lowers the average. Adding a lot of persons with a well-being slightly above 0 might result in a stronger decrease of the welfare function compared to adding one individual with a small *negative* lifetime well-being. In other words, adding one person whose life is not worth living might be better than adding thousands of people whose lives are (barely) worth living.

This seems counter-intuitive, but I believe we are dealing with a moral illusion here. As with the mere addition paradox, three remarks are in order to sufficiently weaken the weak sadistic conclusion.

First of all, both additions decrease the welfare function, and it is possible to avoid such decreases by simply not adding any of those people. So it is not a dilemma between adding the one sufferer versus adding the barely happy people, but a trilemma between those two options and a third option: no addition. The latter option is always preferable.

Second, the weak sadistic conclusion only occurs when the mentioned well-being levels of the added people are as high as they can possibly get. In more realistic cases, the well-being levels of the added people are not maximal. For example, in the above scenarios moving to the repugnant conclusion, we assumed that we could redistribute well-being from $B=(100;98)$ to $C=(99,99)$. So the well-being 98 of the added person is not its maximal possible level. If the well-being was already maximal, then the repugnant conclusion is already avoided because a move to situation C would not be possible. Similarly, if the well-being levels of the added people in the problem of the sadistic conclusion are not fixed or maximal, a redistribution of well-being is possible. And the sacrifice for the better-off people is low when they have to redistribute their well-being with the one person at a negative level. In contrast, when the better-off people have to redistribute their well-being with the thousands of people at a low positive level of well-being, their well-being might drop drastically towards a very low averaged level.

Related to this is the abovementioned suggestion to write the welfare function as a sum of two parts $W = W_{ch} + W_{add}$. If the added person with a negative well-being belongs to the changeable group, i.e. if other people can transfer their well-being to this miserable person, the weak sadistic conclusion becomes weak, because those other people would prefer a small redistribution of their well-being towards the one miserable person over a huge redistribution of their well-being towards a huge number of added people who have small positive levels of well-being. But if a transfer of well-being towards the added

person is impossible, the added person belongs to W_{add} which contains a sum of well-being. Adding a miserable person with negative well-being would lower W_{add} .

Third, as mentioned above, we can say that procreation is allowed as a deontic permission, as long as the new lives are worth living.

4.6.7.3 No replication invariance

The above welfare function contains a positive and a negative part. Those two parts are weighted with two different population factors $P_{s,N_R}(N_+)$ and $P_{t,N_R}(N_-)$. This implies that the theory is not replication invariant when a situation X is replicated into a situation $X' = X + X$, doubling the population size. The welfare function of this doubled situation X' is not simply proportional to the welfare function of the old situation: $W(X')$ does not equal for example $2W(X)$. This means that reversals might occur after replication: if $W(X) > W(Y)$ it can happen that $W(X') < W(Y')$. Replication invariance is defined as the impossibility of such reversals.

It is easy to demonstrate that our welfare function has no replication invariance. There are in fact two reasons why there is no replication invariance. First, suppose that there is no negative part in the welfare function. In that case, the theory is replication invariant only when the parameter $s=0$. When the parameter $s>0$, we would still have a violation of replication invariance (unless we would also duplicate N_R into $N'_R = 2N_R$, but that is cheating). But this is a rather weak violation of replication invariance, because it only occurs for intermediate population sizes. When N is very low or very high, the theory becomes replication invariant.

A second, more serious violation of replication invariance occurs for mixed populations when the two population factors are different.³³ We could restore replication invariance for such mixed populations by adapting the welfare function into for example:

$$W(x) = P_{s,N_R}(N)\langle x^+ \rangle_p - P_{s,N_R}(N)\langle |x^-| \rangle_q.$$

The population factors are now the same for the positive and negative parts, and they use the total number of beings N .

This welfare function has replication invariance for mixed populations, but unfortunately it implies the strong sadistic conclusion: in some situations the welfare function might increase by adding a person with a (small) negative lifetime well-being. Yet, this strong sadistic conclusion can be avoided by introducing another deontological constraint that prohibits the addition of a miserable person to increase the welfare function. This deontological constraint refers to the mere means principle that will be

³³ This kind of violation of replication invariance is equivalent to Parfit's 'absurd conclusion' (Parfit, 1984, p. 410).

discussed in section 6.2. The miserable person would be used as merely a means to increase the welfare function, because the person has to undergo and do something that s/he does not want: s/he has to be born and live, whereas s/he would rather not live at all. Such deontological mere means principle helps to avoid the strong sadistic conclusion.

Nevertheless, when $s > 0$, we are still stuck with a violation of replication invariance. There does not seem to be a solution, an adaptation of the welfare function, that respects replication invariance without violating another important moral intuition. The question is: which intuition is the weakest? I would answer that the violation of replication invariance is less bad than a violation of e.g. the (reverse) repugnant conclusion or the strong sadistic conclusion.³⁴

In summary, the mere addition paradox, the weak sadistic conclusion and the lack of replication invariance are three weak counter-intuitive implications of the theory. We can simply split the welfare function into changeable and additional parts and we can furthermore add a deontic permission: procreation is always allowed (as long as the maximum attainable level of well-being of the added person is positive). This is sufficient to deal with those problems.

4.6.8 Intermezzo: a more complex formulation to solve the replaceability problem

The above welfare function uses the levels of lifetime well-being as input parameters. Hence, an impartial observer behind the veil of ignorance is required to group momentaneous minds together into mutually exclusive subsets that represent strictly separated persons. However, as we have seen in the section on personal identity and psychological continuity, personal identity between momentaneous minds is not always an all-or-nothing issue.

In this intermezzo I want to propose a more accurate welfare function that deals with a more complex account on personal identity. All we have is a set of momentaneous

³⁴ Note that a generalized f-mean (Kolmogorov, 1930)

$$W_{f_a}(x(h)) = f_a^{-1} \left(\frac{1}{N_F} \sum_{i=1}^{N_F} f_a(x_i) \right).$$

that includes the negative levels of well-being and does not include a population factor, has replication invariance. But this theory requires a deontological constraint (the mere means principle discussed in section 6.2) and a deontological permission (the 3-N-principle discussed in section 10.4) to avoid respectively the strong sadistic conclusion and the reverse repugnant conclusion.

minds (indexed by $\pi(t)$, i.e. momentaneous person π at time t) who experience a momentaneous well-being $\mu_{\pi(t)}$. The impartial observer behind the veil does not have to group these momentaneous minds into an individual's lifetime well-being x_i . Hence, difficulties in grouping momentaneous minds into subsets that represent all-or-nothing personal identities can be avoided. Instead, the impartial observer can work with a connectivity function $c_{\pi(t),\pi'(t')}$. This connectivity can have two different interpretations that are coherent with each other.

First, behind the veil of ignorance, the connectivity function might be proportional to the conditional probability: if the impartial observer would experience momentaneous well-being $\mu_{\pi(t)}$, then his/her probability to experience momentaneous well-being $\mu_{\pi'(t')}$ (of another momentaneous person π' at another time t') will be proportional to the connectivity. The impartial observer, once incarnated in front of the veil, might travel around between momentaneous minds and have multiple momentaneous experiences belonging to multiple minds. Hence, the idea is that the impartial observer behind the veil first calculates his/her probability to experience momentaneous well-being $\mu_{\pi(t)}$ and next ascribes conditional probabilities to experience other momentaneous minds given that s/he already experienced (or will experience) $\mu_{\pi(t)}$ somewhere during his/her stay in front of the veil.

As a second interpretation, the connectivity function represents how strong the two momentaneous minds $\pi(t)$ and $\pi'(t')$ belong to the same person over time. The function represents the psychological and physical connectivity between two momentaneous minds. The futuristic Parfitian thought experiments (e.g. teleportation, mind copying, mind swapping, splitting minds or changing personalities) imply that momentaneous minds can be mutually related in degrees: it is not an all-or-nothing question whether or not two momentaneous minds belong to the same personal identity over time.

If for example the personal identity splits like the splitting of a rope into two branches (i.e. the momentaneous minds generate a Y-shape in space-time), we can look at three momentaneous minds at times t , $t' > t$ and $t'' > t$: $\pi(t)$ and $\pi'(t')$ are connected through $c_{\pi(t),\pi'(t')} > 0$, and also $\pi(t)$ and $\pi''(t'')$ are connected through $c_{\pi(t),\pi''(t'')} > 0$, but $c_{\pi'(t'),\pi''(t'')} = 0$, and hence $\pi'(t')$ and $\pi''(t'')$ do not belong to the same person although $\pi(t)$ and $\pi'(t')$ do and $\pi(t)$ and $\pi''(t'')$ do.

Now we have to derive the welfare function over a considered time-interval Δt . First, we calculate the time-average number of momentaneous minds

$$v_F = \frac{1}{\Delta t} \int_0^{\Delta t} N_F(t) dt,$$

with $N_F(t)$ the number of momentaneous minds in front of the veil at time t . If the impartial observer will experience $\mu_{\pi(t)}$, then his/her integrated well-being can be written as

$$\hat{\mu}_{\pi(t)} = \int_0^{\Delta t} \sum_{\pi'(t')=1}^{N_F(t')} c_{\pi(t),\pi'(t')} \mu_{\pi'(t')} dt'.$$

As with the values of life x_i^+ and x_i^- , these integrated well-being levels now have to be projected to the positive and negative values $\hat{\mu}_{\pi(t)}^+$ and $\hat{\mu}_{\pi(t)}^-$. Write v_+ and v_- as the time-average number of positive and negative values of integrated well-being, and v_R as a reference number. The welfare function can then be written as a summation (integral) over the momentaneous minds $\pi(t)$:

$$\begin{aligned} W(c, \mu) &= \frac{v_+}{v_+ + v_R} \left(\frac{1}{v_F \Delta t} \int_0^{\Delta t} \sum_{\pi(t)=1}^{N_F(t)} \left(\left(\int_0^{\Delta t} \sum_{\pi'(t')=1}^{N_F(t')} c_{\pi(t),\pi'(t')} \mu_{\pi'(t')} dt' \right)^+ \right)^p dt \right)^{\frac{1}{p}} \\ &+ \frac{v_-}{v_R} \frac{1}{v_F \Delta t} \int_0^{\Delta t} \sum_{\pi(t)=1}^{N_F(t)} \left(\int_0^{\Delta t} \sum_{\pi'(t')=1}^{N_F(t')} c_{\pi(t),\pi'(t')} \mu_{\pi'(t')} dt' \right)^- dt. \end{aligned}$$

This expression does not require a grouping in subsets that represent different beings with a unique personal identity over time. The impartial observer merely has to ascribe values to the levels of connectivity between each two momentaneous minds. This is the connectivity function. The higher the connectivity, the more the impartial observer believes that two momentaneous minds belong to the same person over time.

The connectivity function is new in a consequentialist welfare ethic. It indicates that not only momentaneous well-being matters, but also connections between momentaneous minds matter. The connectivity function can be represented as a web where the momentaneous minds are the nodes and the threads connect the different momentaneous minds. The thicker the thread, the more the two minds are psychologically and physically connected, the more they can be said to belong to the same person over time, and the higher the connectivity function will be.

Now the problem of replaceability can be understood in a simple way: if you kill a person (who has a strong identity over time) and replace him by another person whose momentaneous minds are not connected to the minds of the killed person, it is as if you cut some threads in the connectivity web. Some connectivities are set to zero. Hence, the welfare function decreases. In other words: not only the levels of well-being at the nodes of the web (the momentaneous minds) have moral (intrinsic) value, also the threads between the nodes have moral value.

Let's study the consequences of different connectivity functions. First, the most trivial connectivity function is $c_{\pi(t),\pi'(t')} = \delta_{\pi(t),\pi'(t')}$, i.e. the connectivity is infinite if

$\pi(t) = \pi'(t')$ and 0 otherwise.³⁵ There is no connectivity between different momentaneous minds; there is no personal identity over time. In this case, the different momentaneous minds are treated as completely different persons. It is as if all persons briefly pop up into existence. All persons immediately die and are replaced by other persons. In this case, the problem of replaceability will not be solved: momentaneous minds are fully replaceable, because they are in fact replaced all the time.

The other extreme is to take $c_{\pi(t),\pi'(t')} = 1$, i.e. all momentaneous minds at all times are connected, as if there was only one superperson who experiences everything. In this case, we end up with sum-utilitarianism, simply adding up the momentaneous well-being of all minds. If the connectivity function remains equal to 1 even when persons are replaced, the problem of replaceability will not get solved: replacing a person will not influence the welfare function.

In between the above two extreme options for the connectivity function, there is an interesting one. Suppose the momentaneous minds can be easily grouped into sets that correspond with persons having a clear personal identity over time. We can write the connectivity as

$$c_{\pi(t),\pi'(t')} = \frac{R_i}{T_i + R_i} \frac{Id_i(\pi(t), \pi'(t'))}{N(T_i)}$$

with $Id_i(\pi(t), \pi'(t')) = 1$ if the two momentaneous minds $\pi(t)$ and $\pi'(t')$ belong to the same person i , and it is 0 otherwise. $N(T)$ is a normalization factor that can be included in order to avoid a kind of double counting (the welfare function has a double integral, so the well-being of some momentaneous minds will be counted multiple times). For simplicity we can take $N(T) = T$.³⁶ The factor $R_i/(T_i + R_i)$ might correspond with the impartial observer's probability to experience momentaneous mind $\pi'(t')$ given the experience of $\pi(t)$. In another interpretation, the parameter R_i is a reference time-length of individual i and T_i is the lifespan of that individual. We briefly encountered this expression in the section on the lifetime perspective (section 4.2.4). As was mentioned there, the reference time-length was related to the psychological connectivity. The above expression of the connectivity function clearly demonstrates this relation. After plugging the connectivity function in the welfare function and setting $N(T) = T$, we get a simplified expression:

$$W(T, \bar{\mu}) = \frac{N_+}{N_+ + N_R} \left(\frac{1}{N_F} \sum_{i=1}^{N_F} \left(\frac{T_i R_i \bar{\mu}_i^+}{T_i + R_i} \right)^p \right)^{\frac{1}{p}} + \frac{N_-}{N} \frac{1}{N_F} \sum_{i=1}^{N_F} \frac{T_i R_i \bar{\mu}_i^-}{T_i + R_i},$$

³⁵ Mathematically speaking, this is the Dirac delta function.

³⁶ Due to this normalization factor, the connectivity function approaches the Dirac delta function when $T \rightarrow 0$.

with $\bar{\mu}_i^+$ the time-average momentaneous well-being of individual i , averaged over the lifespan T_i , and projected on the positive values. We encountered this welfare function before, when the value of life equals

$$x_i^+ = \frac{T_i R_i \bar{\mu}_i^+}{T_i + R_i}.$$

As mentioned before, this expression avoids the problem of replaceability. It also presents a new solution to the excessiveness (demandingness) problem of prioritarianism raised by Holtug (2007): the animals with the shortest lives are the worst-off and should get the highest priority. As humans have a long lifespan, humans should sacrifice a lot in order to increase the lifetime well-being of animals such as frogs.

This prioritarianism is very demanding for humans. To weaken the demandingness objection, Holtug (2007) proposed a time-slice prioritarianism: at each separate moment of time, momentaneous well-being should be distributed in a prioritarian way. This is less demanding for humans than a lifetime perspective prioritarianism, because as we have seen above, the gap in momentaneous well-being between a human and a frog is lower than the gap in lifetime well-being (the difference in lifespan and the difference in psychological connectivity between humans and frogs add to the gap in lifetime-well-being).

But Holtug also mentioned some counter-intuitive problems of his proposal. His time-slice prioritarianism is insensitive to inter-temporal compensations that can take place in a life. Suppose individual A has a high lifetime well-being, but at time t she has a very low momentaneous well-being. In contrast, individual B has a high momentaneous well-being at time t but a low lifetime well-being. Then time-slice prioritarianism says that at time t well-being should be redistributed from B to A. But if this is the only moment when a redistribution between A and B is possible, A ends up with an even higher lifetime well-being.

In contrast, the above presented connectivist prioritarianism not only has a weaker demandingness objection just as Holtug's proposal, but it avoids the counter-intuitive conclusion of time-slice prioritarianism as well. With the above expression, if the animal has a low connectivity and hence a low value R_i , a unit increase in average well-being or in lifespan will give a relatively low increase in lifetime well-being, compared with a unit increase in average well-being or in lifespan for a person having a strong connectivity (a high value R_i). This will compensate the priority for the shortest lifespan, resulting in a less demanding ethic for humans.

We can also look at the issue of abortion. As McMahan (2002) argued, killing an adult human is worse than killing an early human embryo, because the human has a stronger connectedness with his future self than the embryo has with hers. For the embryo, death means the loss of a higher amount of future well-being, because the embryo has a

longer potential future ahead of her compared to the adult. But this amount of future well-being should be multiplied by the psychological connectedness that the embryo has with her future self at the moment she is killed. As this psychological connectedness is very low, the result is a lower loss of weighted future well-being for the embryo.

The above can be mathematically expressed with the integrated well-being $\hat{\mu}_{\pi(t)}$, which includes the connectivity function. This means that the difference between $\hat{\mu}_{\pi(t)}$ (the integrated well-being of the non-killed embryo π at time t) and $\hat{\mu}_{\pi(t)}^{\dagger}$ (the integrated well-being of the embryo when she is killed at time t) is lower than the difference between $\hat{\mu}_{\pi'(t')}$ (the integrated well-being of the non-killed human adult π' at time t') and $\hat{\mu}_{\pi'(t')}^{\dagger}$ (the integrated well-being of the adult when he is killed at time t').

Even if this difference of the integrated well-being is lower for the embryo, it does not yet imply that the difference of the welfare function between killing and not killing the embryo is lower compared to killing and not killing the adult. That is because killing an embryo not only results in a slightly lower integrated well-being $\hat{\mu}_{\pi(t)}^{\dagger}$, but it also results in the loss of all future integrated well-being states. The latter loss can easily be bigger than the loss of all future integrated well-being states of the adult.

In this intermezzo I argued that not only the momentaneous experienced well-being is important in a consequentialist welfare ethic. The momentaneous well-being is experienced by momentaneous minds, and these minds form the nodes of a vast connectivity web. Between two nodes can be a link: the connectivity that measures how strongly the two momentaneous minds are psychologically connected. Not only has each node a moral value, the momentaneous well-being, but also each link has a moral value, the connectivity. Lowering the value of a node lowers the welfare function, but also lowering the connectivity between nodes lowers the welfare function.

With this connectivist welfare ethic, we get a new, elegant reframing of the replaceability problem: killing and replacing persons that have the same momentaneous well-being means cutting links between nodes, setting some connectivities to zero. The number and values of the nodes remain the same, but the connectivity web is less connected. Hence, the welfare function decreases even if the momentaneous well-being remains the same.

The connectivity web also allows for intrapersonal, intertemporal shifts in well-being: well-being can be shifted from node to node, following a connected path between the nodes. The connectivity allows for those shifts, because connectivity means that the two nodes belong to the same person. However, if the connectivity between two nodes is very weak, not much well-being is allowed to be shifted from one node to the other. If there is a weak connectivity between my current self and my future self over 30 years, my future self should be treated as an almost different person. As a consequence,

smoking today will not only be imprudent but also to some degree immoral: it harms (without permission) an almost different future person who will get cancer.

Finally, the connectivity is not necessarily a binary function that takes only two values: one (personal identity between two momentaneous minds) and zero (no identity). This means that it is not an all-or-nothing issue whether two momentaneous minds belong to the same person: it can be a matter of degree. Personal identity over time can be more fluid, especially if futuristic thought experiments such as teleportation, mind copying, mind splitting and mind swapping would one day become reality. We now already have an ethical theory that is fit to deal with those tricky situations.

What about fractional consciousness?

My intuition says that consciousness is an all-or-nothing issue: either it is switched on or it is switched off, either the neural program runs correctly or it doesn't. Still one might object that scientists might discover that consciousness comes in matters of degree. We should distinguish quality versus quantity of an experience. A fractional consciousness deals with the quantity of an experience, not the quality. The quality can have e.g. an intensity, and intensity of feelings can come in degrees. For example pain is stronger than another. This difference in intensity is a difference in the quality of an experience: a weak pain feels different from a strong pain. On the other hand, fractional consciousness deals with the quantity of an experience and asks the question: what if there is not an integer, but a fractional number of momentaneous minds? What if consciousness itself is a matter of degree instead of an on/off switch? Bostrom (2006) offers the most challenging thought experiments to argue that at least in theory, fractional minds are conceivable. Bostrom's thought experiments extend those of Parfit (1984) and deal with brain-duplication to argue that consciousness comes in degrees, that there might be for example 1,5 instead of two conscious people.

The above welfare function can be easily adapted to deal with fractional consciousness. First, we replace the integer number of momentaneous minds at time t by:

$$N_F(t) = \sum_1^{N_F(t)} 1 \rightarrow \sum_{\pi(t)=1}^{N_F(t)} \alpha_{\pi(t)},$$

with $\alpha_{\pi(t)}$ a fractional number for a momentaneous mind $\mu_{\pi(t)}$. This fractional number can be e.g. 0,75 or 1,32 or whatever. It corresponds with the fractional number of minds that have the same mental state, i.e. the mental state of the momentaneous mind $\mu_{\pi(t)}$. In the numerator of the welfare function, we replace $\mu_{\pi(t)} \rightarrow \alpha_{\pi(t)}\mu_{\pi(t)}$. If the fractional number is 1, we get the previous welfare function. If a momentaneous mind comes with a very low level consciousness, the fractional number is so low that the momentaneous mind does not strongly influence the welfare function.

4.6.9 Uncertainty aversion

In the previous sections, we assumed a uniform probability distribution behind the veil of ignorance: the impartial observer knows that s/he will be born as individual i with probability $1/N$. This assumption can be justified: it is the only probability distribution with maximal information entropy, i.e. least information content (Cover & Thomas, 1991, chapter 11), at least if you have no further information about, for example, what the average well-being or the standard deviation will be.

However, we can also assume that behind the veil, you do not know the probability distribution, and you have an uncertainty aversion.³⁷ This is what Rawls originally intended in his theory of justice (Rawls, 1971). A person who has a maximum (unrestricted) uncertainty aversion prefers playing any gamble with known probabilities above a gamble with unknown probabilities. Any gamble also means: the gamble with the worst probability. The maximum uncertainty averse person would prefer playing the gamble where s/he will become the worst-off person for sure, instead of playing the gamble where s/he does not know the probability to become the worst-off person. As mentioned above, the weights a_i refer to the probability to get a lifetime well-being level $x_{[i]}$. Hence, the worst probability distribution is the one with $a_1=1$, i.e. the highest probability to become the worst-off individual. As with the situation of maximum risk aversion, maximum (i.e. unrestricted) uncertainty aversion also results in maximin (see Gilboa & Schmeidler, 1989).

Gajdos & Kandil (2008) gave a mathematical proof for a welfare function that is a linear combination of maximin and average utilitarianism³⁸:

$$W_{GK} = \theta x_{[1]} + (1 - \theta) \langle x \rangle_1,$$

where $x_{[1]}$ is the worst-off and θ is a parameter between 0 and 1. In other words: $a_1 = 1/(1+(N-1)(1-\theta))$ and $a_{i>1} = (1-\theta)/(1+(N-1)(1-\theta))$. The proof is based on a restricted kind of uncertainty aversion (restricted mixture neutrality). No uncertainty aversion corresponds with $\theta=0$, unrestricted (maximum) uncertainty aversion corresponds with $\theta=1$.

The restricted mixture neutrality used in the proof, is only one way to restrict uncertainty aversion. Although I do not prove it here, I postulate that there are other restrictions of uncertainty aversion that will result in other weights a_i , and perhaps one kind of restriction results in a moderate egalitarianism (Jensen, 2003) discussed below.

³⁷ Risk aversion assumed a knowledge about the probability distribution, and resulted in a concave utility function $\notin \mathcal{X}^p$. Uncertainty aversion assumes a lack of knowledge of the probability distribution.

³⁸ Gajdos & Kandil assumed a risk neutral but uncertainty averse impartial observer.

4.6.9.1 Moderate egalitarianism

Like Gajdos & Kandil (2008), I choose the weights in such a way that we get average utilitarianism and maximin as limits. Take for example:

$$a_l = N \frac{Q^{l-1}}{\sum_{j=1}^N Q^{j-1}},$$

with $Q \in [0,1]$. If $Q=0$, we get maximin; if $Q=1$ (and the power $p=1$, i.e. no risk aversion), we get average utilitarianism. A welfare function with the above weights is called a generalized Gini welfare function (Weymark, 1981). When distributing benefits, the worst-off should get a priority because they have a higher weight factor a_l . In contrast with (power mean) prioritarianism, the priority in this moderate egalitarianism depends on the position of an individual *relative to the well-being levels of the others*. Prioritarianism uses a concave utility function $u(x)$ that does not depend on the relative position (relative to the well-being of others).

This generalized Gini welfare function can be derived as follows. First, start with a population of two individuals and two situations: $X = (x_{[1]}, x_{[2]})$, and $Y = (y_{[1]}, y_{[2]})$. We can write three consequentialist theories of justice in a simple set of mathematical inequalities:

- Strict egalitarianism: X is better than Y if and only if

$$x_{[2]} - x_{[1]} \leq y_{[2]} - y_{[1]}$$

i.e. the difference between the values of life should be minimized.

- Maximin: X is better than Y if and only if

$$x_{[1]} \geq y_{[1]}$$

i.e. the value of life of the person in the worst position should be maximized.

- Sum-utilitarianism: X is better than Y if and only if

$$x_{[1]} + x_{[2]} \geq y_{[1]} + y_{[2]}$$

i.e. the total value of life (total utility) should be maximized.

These expressions can be unified in one inequality

$$x_{[1]} + Qx_{[2]} \geq y_{[1]} + Qy_{[2]}$$

where the parameter Q takes the values:

- $Q = -1$: strict egalitarianism,
- $Q = 0$: maximin,
- $Q = +1$: utilitarianism.

We can write $W(X) = x_{[1]} + Qx_{[2]}$ as the welfare function of situation X . Generalizing to situations with N number of individuals can be done using a recursive relation. The welfare function of situation X then reads:

$$W(X = (x_{[1]}, \dots, x_{[N]})) = x_{[1]} + Q(x_{[2]} + Q(x_{[3]} + \dots)) = \sum_{i=1}^N Q^{i-1} x_{[i]}.$$

In order to avoid the repugnant conclusion, we have to normalize this welfare function with the sum $\sum_{j=1}^N Q^{j-1}$. The result is the Gini welfare function

$$W_{Gini}(x_{\uparrow}) = \frac{\sum_{i=1}^N Q^{i-1} x_{[i]}}{\sum_{j=1}^N Q^{j-1}}.$$

4.6.9.2 Replication invariance

The weight Q can be adapted to make the theory replication invariant. If we start with a world with two individuals and $x_{\uparrow} = (x_{[1]}, x_{[2]})$, then replication invariance means that making a copy of this world, i.e. $x_{\uparrow}' = (x_{[1]}, x_{[1]}, x_{[2]}, x_{[2]})$, does not change the welfare function. Using the above weights, the generalized Gini welfare function does not remain invariant under replication (see e.g. Fleurbaey et al. 2009). More generally, start with the N -individual situation $x_{\uparrow} = (x_{[1]}, \dots, x_{[i]}, \dots, x_{[N]})$ and replicate this K times $x_{\uparrow}' = (x'_{[1]}, \dots, x'_{[j]}, \dots, x'_{[KN]})$, with $x_{[1]} = x'_{[1]} = x'_{[2]} = \dots = x'_{[K]}$, $x_{[2]} = x'_{[K+1]} = \dots = x'_{[2K]}$ and so on for the other $x_{[i]}$. Applying the welfare function gives

$$W_{Gini}(x_{\uparrow}') = \frac{\sum_{j=1}^N Q^{j-1} x'_{[j]}}{\sum_{k=1}^N Q^{k-1}} = \left(\frac{1 - Q^K}{1 - Q^{KN}} \right) \sum_{l=1}^N Q^{K(l-1)} x_{[l]} = \frac{\sum_{l=1}^N Q^{K(l-1)} x_{[l]}}{\sum_{k=1}^N Q^{K(k-1)}}.$$

This expression is not the same as the above $W_{Gini}(x_{\uparrow})$, but there is a similarity. If we rescale $Q^K \rightarrow Q$, we get the same expression. That means that invariance under replication is restored if the parameter Q depends on the number of individuals. So we can start with a fixed parameter in the two-individuals world: Q_2 . The parameter in a world with N individuals then becomes $Q_N = Q_2^{2/N}$. This is in line with our intuition: the more individuals, the higher the parameter Q_N should become, because otherwise the higher well-being levels will rapidly get extremely low weight factors. So the replication invariant expression for the welfare function of Gini-moderate egalitarianism becomes:

$$W_{Gini}(x_{\uparrow}(h)) = \frac{\sum_{l=1}^N Q_2^{\frac{2}{N}(l-1)} x_{[l]}(h)}{\sum_{k=1}^N Q_2^{\frac{2}{N}(k-1)}}.$$

This shows that a geometric generalized Gini welfare function can be made replication invariant, if the weights properly depend on the population size N .³⁹

³⁹ As with power mean prioritarianism, we can adapt this expression to deal with non-trivial psychological connectivities. Write $\hat{\mu}_{[\pi(t)]} = \left(\int_0^{\Delta t} \sum_{\pi'(t')=1}^{N_F(t')} c_{\pi(t), \pi'(t')} \mu_{\pi'(t')} dt' \right)_{\uparrow}$ in increasing order, and $[\pi(t)] = \tau \in$

4.6.9.3 Avoidance of intransitivity and misery for the ultra rich

The moderate egalitarian welfare function solves Temkin's intransitivity problem (Temkin, 1987). This intransitivity problem comes down to a line of reasoning where choice A is better than B, and B better than C, but C is better than A (as in the rock-paper-scissors hand game). In his argument, Temkin refers to a First Standard View (FSV) and a Second Standard View (SSV).

The FSV states that a situation in which one person suffers a lot is better than a situation where two people suffer a lot, but their suffering is a little bit less than the suffering of the one person in the first situation. In other words: two almost extreme sufferers is worse than one extreme sufferer. And three almost almost extreme sufferers is worse than two almost extreme sufferers. And four almost almost almost extreme sufferers is worse still. Continuing in this way, we end up with a situation where a very high number of people suffer only a little bit (from a light headache). If transitivity applies, then the FSV means that this final situation is the worst of all.

But that conclusion violates another intuition, the SSV, which says that the first situation, where everyone is really happy except one extreme sufferer, is much worse than the final situation where a very high number of people are almost really happy and no-one really suffers.

This SSV corresponds with the problem of the 'Misery for the Ultra-Rich'. One critique of sum-utilitarianism and (power mean) prioritarianism is that according to these theories, there are situations where it is good to sacrifice the well-being of a worst-off person if this sacrifice results in a huge benefit that can be distributed amongst a huge amount of best-off people (the ultra rich), each getting a tiny share that increases well-being. As long as the total number of ultra-rich beneficiaries is high enough, any amount of cost for the worst-off will be offset by the sum of small benefits for the many ultra rich. The ultra rich can get richer while the poorest person gets

$[0, v_F \Delta t]$ with $\int_0^{\Delta t} N_F(t) dt = v_F \Delta t$. Then we can write $W = \frac{\int_0^{v_F \Delta t} Q^{\frac{\tau}{v_F \Delta t}} \hat{\mu}_\tau d\tau}{\int_0^{v_F \Delta t} Q^{\frac{\tau}{v_F \Delta t}} d\tau}$. When

$c_{\pi(t), \pi'(t')} = \frac{R_i}{T_i + R_i} Id_i(\pi(t), \pi'(t'))$, the function $\hat{\mu}_{[\pi(t)]}$ becomes an increasing step function between 0 and $v_F \Delta t$ where each step corresponds with the life of an individual (and the width of a step, T_i corresponds with

the lifespan T_i such that $\sum_{i=1}^{N_F} T_i = v_F \Delta t$). The welfare function then reduces to $W = \frac{\sum_{i=1}^{N_F} Q^{\frac{\sum_{j < i} T_j}{v_F \Delta t}} (1 - Q^{\frac{T_i}{v_F \Delta t}}) x_{[i]}}{(1 - Q)}$

with $x_{[i]} = \frac{T_i R_i \bar{\mu}_i}{T_i + R_i}$. This is a slight generalization of the Gini welfare function: only when all $T_i = T_j$, i.e. all

lifespans are equal, we get $W_{Gini}(x_\uparrow) = \frac{\sum_{i=1}^{N_F} Q^{\frac{1}{N_F}(i-1)} x_{[i]}}{\sum_{k=1}^{N_F} Q^{\frac{1}{N_F}(k-1)}}$.

poorer. This seems counterintuitive according to many people (Broome, 2004 p.58; Holtug, 2006 p.134; Dorsey, 2009, p.54).

Let's apply the intransitivity problem to the equations of moderate egalitarianism with the Gini welfare function. We first note that in contrast with the repugnant conclusion, in the intransitivity problem the total number of people N_Z is constant and very large. Let's start with history h_1 where N_1 (much smaller than N_Z) of extreme sufferers have a well-being w below the happy state H , and all the other $N_Z - N_1$ people are in the happy state H . So we get:

$$W_{Gini}(h_1) = \frac{\sum_{i=1}^{N_1} Q^{i-1}(H - w) + \sum_{j=N_1+1}^{N_Z} Q^{j-1}H}{\sum_{k=1}^{N_Z} Q^{k-1}},$$

with

$$Q = Q_{N_Z} = Q_2^{2/N_Z}.$$

In history h_2 there are $N_2 > N_1$ people with well-being $H - w + \delta_1$ (with δ_1 small but positive), and the others still have well-being H . For history h_3 there are $N_3 > N_2$ people with well-being $H - w + \delta_1 + \delta_2$. Moving on, we get history h_p where N_p people have well-being $H - w + \Delta_p$, with $\sum_{k=1}^{p-1} \delta_k = \Delta_p$.

Let's suppose that the FSV is always valid. Then from the inequality $W_{Gini}(h_1) \geq W_{Gini}(h_p)$ and using

$$\sum_{i=1}^N Q^{i-1} = \frac{1 - Q^N}{1 - Q},$$

we get

$$(H - w) \left(\frac{1 - Q^{N_1}}{1 - Q} \right) + H \left(\frac{1 - Q^{N_p}}{1 - Q} - \frac{1 - Q^{N_1}}{1 - Q} \right) \geq (H - w + \Delta_p) \left(\frac{1 - Q^{N_p}}{1 - Q} \right).$$

When N_p becomes very large, and Q is lower than 1, the factor Q^{N_p} goes to zero. This simplifies the above inequality to: $\Delta_p \leq wQ^{N_1} < w$.

We can conclude that if the FSV is always satisfied, the total increment Δ_p is always *strictly* lower than w . So we will never be able to get the well-being arbitrarily close to H . If we want to proceed and move close to H , we at one point have to violate the FSV. So we have an history h_f where the inequality flips:

$W(h_1) > W(h_2) > \dots > W(h_f) < W(h_{f+1}) < \dots$. Only when $Q = 1$ (sum-utilitarianism) we have the constraint $\Delta_p \leq w$ that allows us to move close to H . And if $Q = 0$, then $\Delta_p \leq 0$ and FSV is always violated. Moderate egalitarianism solves Temkin's intransitivity paradox, in the sense that it has a point in the series where the FSV no longer becomes valid and the SSV takes over.

4.6.9.4 The problem of independence and Allais paradox

As we have seen, unweighted power mean prioritarianism has a welfare function that has the property of independence or strong separability (see e.g. McCarthy, 2008), except for mixed populations and variable populations. The situation with moderate egalitarianism is worse: strong separability is violated, even when populations are not mixed and population size is constant.

Suppose we have two times two histories, each involving three individuals. History $h_1=(1;1;1)$, $h_2=(1;0;5)$, $h'_1=(0;1;1)$ and $h'_2=(0;0;5)$. We see that history h'_1 is related to history h_1 , just as history h'_2 is related to h_2 . Suppose the impartial observer behind the veil can choose between either h_1 and h_2 or between h'_1 and h'_2 . Independence now says that h'_1 is better than h'_2 if and only if h_1 is better than h_2 . The presence of the first person should not matter (whether s/he has well-being 1 in the first two histories, or well-being 0 in the last two histories). We can apply the welfare function W_{Gini} . When $Q < 1/2$, we get $W_{Gini}(h_1) = 1 + Q + Q^2 > W_{Gini}(h_2) = Q + 5Q^2$. But if $Q > 1/4$, we see that $W_{Gini}(h'_1) = Q + Q^2 < W_{Gini}(h'_2) = 5Q^2$. In other words, when Q is in the range $\frac{1}{4} < Q < \frac{1}{2}$, it is possible that situation h'_2 is better than situation h'_1 , although situation h_2 was worse than situation h_1 . Independence is restored when we exclude the first person.

In decision theory, Allais paradox (Allais, 1953; Kahneman and Tversky, 1979) is similar to the problem of independence. An example of the paradox goes as follows. Suppose we have two experiments, each consisting of two gambles. In the first experiment, you can choose between gamble 1, where you have probability 100% to gain \$100, and gamble 2, where you have probability 89% to gain \$100, 1% to gain nothing and 10% to gain \$500. Most people would choose to play gamble 1, especially when they have risk aversion, because in gamble 2 they risk to gain nothing. In gamble 1 they are always certain to receive some benefit.

In a second experiment, people can choose between gambles 1' and 2'. Gamble 1' has the following probabilities: 89% to gain nothing and 11% to gain \$100. Gamble 2' has 90% to gain nothing and 10% to gain \$500. In this second experiment, people would prefer to play gamble 2', because they gain \$500 at an almost equal probability as gaining \$100 in gamble 1'. The curious thing is that both experiments are in fact quite similar, as can be seen by writing it as in the following table.

Gamble 1	Gamble 2	Gamble 1'	Gamble 2'
89% 100\$	89% 100\$	89% 0\$	89% 0\$
1% 100\$	1% 0\$	1% 100\$	1% 0\$
10% 100\$	10% 500\$	10% 100\$	10% 500\$

In the first experiment, we can decouple a 89% probability to gain \$100. So we can write for gamble 1 that you have 89%+1%+10% to gain \$100. But if we would now set the

gains for those 89% in gambles 1 and 2 equal to zero, we get gambles 1' and 2'. So it is strange why people prefer gamble 1 over 2, but 2' over 1', because gamble 2' is quite similar to gamble 2, and 1' is similar to 1. Gambles 1 and 2 are both changed in the very same way (setting a gain to zero), and yet the order of preference of the gambles suddenly changes.

The analogy with this Allais paradox and the above veil of ignorance example is clear.⁴⁰ If the impartial observer behind the veil would have preferences as in Allais paradox, s/he could end up with a welfare function of moderate egalitarianism.

4.6.9.5 Summary

In summary, we basically have two theories that have maximin as a limit: moderate egalitarianism with a generalized Gini welfare function (having maximin in the limit $Q \rightarrow 0$) and prioritarianism with a power mean welfare function (having maximin in the limit when the power $p \rightarrow -\infty$). Moderate egalitarianism solves Temkin's intransitivity problem, but this theory is not strongly separable, not even when population size is fixed. Power mean prioritarianism does not solve Temkin's intransitivity paradox, because it always respects the FSV and hence it faces the problem of the Misery for the Ultra-Rich. If the latter problem is considered less bad than the problem of independence (strong separability) for unmixed, fixed populations (i.e. for situations with constant numbers of people with positive and negative levels of well-being), then prioritarianism is the better theory. Both moderate egalitarianism and power mean prioritarianism can be derived from a veil of ignorance with respectively uncertainty aversion and risk aversion. Both theories can be unified in one expression: a weighted power mean.

4.6.10 Prioritarian theories for lotteries

In the above descriptions, the impartial observer had to choose between different histories from behind the veil of ignorance. However, it is possible that outcomes are not certain. In this case of probabilistic outcomes, we have to apply the theory to lotteries (Rabinowicz, 2002; McCarthy 2003; 2006; 2008; Otsuka & Voorhoeve, 2009).

⁴⁰ In the veil of ignorance example, the impartial observer might ascribe a probability 1/3 for each individual, because we have a population of 3 persons. If we would like to use the probabilities 89%, 1% and 10% as in the former example, we can take a population of 100 individuals, whereby 89 individuals have well-being 1 in the first history, and so forth.

A lottery L can be written as a set of m histories, where each history has a probability and all probabilities sum to one: $L = \{(p_1; h_1), (p_2; h_2), \dots (p_m; h_m)\}$ with $\sum_{r=1}^m p_r = 1$ and h_r the r -th possible history. From behind the veil of ignorance, we now have to choose between different lotteries (instead of different histories). So there are now two elements of risk: first you don't know who you will be, and second you don't know which history will be actualized.

We can write the expected well-being of individual i in lottery L over all histories as:

$$\langle x_i(L) \rangle = \sum_{r=1}^m p_r x_i(h_r).$$

The welfare function $W_+(L)$ can take three different forms: an ex-ante, ex-inter and ex-post (for simplicity I take again a uniform distribution $a_i=1$).

$$\begin{aligned} W_+^{EA}(L) &= \left(\frac{1}{N_F} \sum_{i=1}^{N_F} \langle x_i^+(L) \rangle^p \right)^{\frac{1}{p}} = \left(\frac{1}{N_F} \sum_{i=1}^{N_F} \left(\sum_{r=1}^m p_r x_i^+(h_r) \right)^p \right)^{\frac{1}{p}}, \\ W_+^{EI}(L) &= \left(\left\langle \frac{1}{N_F} \sum_{i=1}^{N_F} x_i^+(L)^p \right\rangle \right)^{\frac{1}{p}} = \left(\sum_{r=1}^m p_r \frac{1}{N_F} \sum_{i=1}^{N_F} (x_i^+(h_r))^p \right)^{\frac{1}{p}}, \\ W_+^{EP}(L) &= \left\langle \left(\frac{1}{N_F} \sum_{i=1}^{N_F} x_i^+(L)^p \right) \right\rangle^{\frac{1}{p}} = \sum_{r=1}^m p_r \left(\frac{1}{N_F} \sum_{i=1}^{N_F} x_i^+(h_r)^p \right)^{\frac{1}{p}} \end{aligned}$$

In the ex-ante version, we first calculate the expectation values of the well-being levels, and afterwards use them in the welfare function, whereas in ex-post we apply the expectation value at the end, after calculating the welfare functions for the different histories. Ex-inter is a so-called weighted power mean, weighted by the probabilities of the lotteries.

Let's apply our three theories of prioritarianism to the following example. Suppose there are four lotteries, each with two persons (A and B) and two histories (h_1 and h_2). The values of well-being are summarized in the following table.

	L_1		L_2		L_3		L_4	
	h_1	h_2	h_1	h_2	h_1	h_2	h_1	h_2
A	2	2	2	0	2	0	1	1
B	0	0	0	2	2	0	1	1

L_1 is the worst lottery, because we know for sure that only A will win (B will get 0), ending up with an inequality. Lottery L_2 is a bit better, because now both A and B at least get an equal chance to win. L_3 is better still, because we will always end up with a situation of equality, and both A and B get an equal probability to win. Finally, from a

risk averse point of view, lottery L_4 is the best, because now each person will at least win something, and both will end up equal.

The following table summarizes the welfare functions for all four lotteries.

	L_1	L_2	L_3	L_4
W_+^{EA}	$2^{1-\frac{1}{p}} < 1$	1	1	1
W_+^{EI}	$2^{1-\frac{1}{p}} < 1$	$2^{1-\frac{1}{p}} < 1$	$2^{1-\frac{1}{p}} < 1$	1
W_+^{EP}	$2^{1-\frac{1}{p}} < 1$	$2^{1-\frac{1}{p}} < 1$	1	1

We see that each of the three welfare functions satisfies one of our judgments.⁴¹ For example according to the ex-ante W^{EA} , the first lottery is the worst, the other three lotteries are equal. However, the ex-ante welfare function encounters a problem when it comes to variable populations, i.e. when the population size N differs between histories. If person B does not exist in history h_2 , we cannot calculate W^{EA} .⁴²

How to decide between the ex-inter and ex-post versions of the welfare function? We have seen that from behind a veil of ignorance, not knowing who you will be, there is some risk involved, and prioritarian theories show a non-zero risk aversion. When the choice is between lotteries, a second kind of risk is introduced. In the above example, there is the risk of becoming person A instead of B (a risk related to impartiality), and there is the risk of having one history instead of the other (a risk related to lotteries). The ex-inter welfare function treats these two risks in a same way: W^{EI} has coherent risk aversion towards both kinds of risks (this can be seen in the mathematical expression: the two probabilities p_r and $1/N_r$ are treated in the same way). The ex-post welfare function, however, treats these two kinds of risks differently.

This apparent incoherence of ex-post prioritarianism is not a real threat to the theory, because one might argue that the two kinds of risk are not comparable and should not be treated in the same way. The impartiality risk can be related to the notion of separability of persons, which is – morally speaking – something other than a separability of outcomes of a lottery. If the separability of persons is morally more relevant than the separability of outcomes, the two kinds of risks can be treated differently.

⁴¹ As with populations ethics (problems related to variable populations), situations of risk also has unavoidable counter-intuitive implications: it is not possible (or very difficult) to respect all of our seemingly self-evident moral intuitions. See e.g. Dougherty (2013) who argued that we have intuitions both for and against the ex-ante view.

⁴² Assuming that person B has well-being 0 in history h_2 would not work, because why should one include this potential person in the calculation and not include all other potential persons as well? Including all potential persons, setting their non-existent well-being levels equal to zero, will result in a division by an infinite population size N .

Furthermore, Otsuka and Voorhoeve (2009) expressed the intuition that a one-person decision under the risk of a lottery should be treated differently than a multi-person decision (without the risk of a lottery). The one-person decision under risk should follow the expected utility theory to reflect the ‘unity of the individual’ (see also Porter, 2012). When there is only one individual, the ex-post welfare function indeed becomes the expected well-being: $\langle (x^p)^{1/p} \rangle = \langle x \rangle$.

The multi-person decision should follow the prioritarian approach to reflect the separability of persons. So, the incoherence of ex-post prioritarianism might not be a weakness after all, as it might reflect the intuitions shared by e.g. Otsuka and Voorhoeve.

The veil of ignorance can also be reframed in order to make it compatible with this apparent incoherence of the difference in risk attitude between situations with multiple persons versus multiple outcomes. If there are N individuals in front of the veil, you are going to live the life of just one of those individuals. You will be person i with a probability $1/N$, and it is not irrational to be risk averse with respect to the lifetime well-being of this one person. On the other hand, if you would ‘reincarnate’ N times and live the lives of all N persons, you will experience everything and there is no reason to be risk averse towards an individual well-being behind the veil of ignorance.

The point is that behind the veil of ignorance, lotteries can be understood as reincarnations. If in a lottery person i has a probability $p=S/T$ to have an outcome O and a probability $q=(T-S)/T$ to have outcome O' (with T and S natural numbers), it is as if you will live the life of person i a number of T times, S of which you will get outcome O , and $T-S$ of which you will get outcome O' . It is as if you will be reincarnated T times into parallel worlds, leading again the life of person i . There is no reason to be risk averse if you will experience everything of the T copies of person i .

The ex-post approach does however have a problem of independence. Compare the above lotteries L_2 and L_3 . In both lotteries, the outcomes of individual A are the same (outcome 2 of history h_1 is realized, outcome 0 if the second history is realized). If there is independence between the two persons, we can change the outcomes of person A in both lotteries in the same way. The following table represents such a transformation into new lotteries L'_2 and L'_3 .

	L'_2		L'_3	
	h_1	h_2	h_1	h_2
A	0	2	0	2
B	0	2	2	0

We saw that $W^{EP}(L_3) > W^{EP}(L_2)$, because L_3 has always equal outcomes for both individuals (either both 2 or both 0). But now we have $W^{EP}(L'_3) < W^{EP}(L'_2)$, so independence is no longer valid in an ex-post theory (see also Fleurbaey and Zuber, 2012).

4.6.11 Democratic impartial preferences of moral agents

The prioritarian theory uses undetermined values: the risk parameter p^{43} , the reference population size N_R , the choice between ex-ante, ex-inter and ex-post decisions, and the values of life (lifetime well-being levels) x_i . As mentioned in a previous section (4.2), the values of life are functions of experienced well-being of individuals, but these experienced well-being levels are not interpersonally comparable, just like I cannot compare my perception of a red color with your perception of red. Therefore, the values of life are the values attributed by an impartial observer behind a veil of ignorance. This impartial observer uses empathy to guess how situation x for person i would compare to situation y for person j .

In reality, we do not have an ideal impartial observer. What we do have, are moral agents. These are all persons who are able to perform the thought experiment, using empathy or imagination. Each moral agent will imagine him/herself as an impartial observer behind the veil.

In a first step, a moral agent a classifies all momentaneous minds in subsets that represent the N_F^a different real world persons: the number of persons in front of the veil, according to moral agent a (see the section on personal identity and continuity). This classification reflects the importance of personal identity and non-replaceability according to moral agent a (see the section on the replaceability problem above). If the moral agent has no problem with replaceability of persons, s/he can treat all momentaneous minds independently as different persons. Alternatively, the moral agent can ascribe a connectivity function between all the momentaneous minds to represent the level of irreplaceability of persons (see the intermezzo).

In a second step, the moral agent a performs the thought experiment, choosing his/her preferred welfare function (e.g. a positive number-dampened power mean prioritarianism with negative total utilitarianism, or an exponential lifetime well-being generalized f-mean), deriving his/her own preferred risk aversion parameter p^a , population reference N_R^a , lottery decision rule and estimates of lifetime well-being x_i^a (or momentaneous well-being μ_π^a and connectivity function c^a). These parameters and values can lie in a certain range: the broader the range of e.g. the risk aversion parameter, the more flexible is moral agent a 's attitude towards risk aversion.

Next, the moral agent calculates his/her own welfare function $W^a(p^a, N_R^a, x_i^a)$ with respect to the range of parameters. Of course, different moral agents will end up with different welfare functions and different maximizations, even when they do the thought experiment as sincerely as possible. Who is right? Who has the best risk attitude? Who

⁴³ Or $\ln(a)$ in case of the exponential lifetime well-being in the Kolmogorov mean.

has the best empathy? We have to respect the principle of universalization: if you are allowed to do the exercise (the thought experiment of the veil of ignorance) using your preferences (e.g. your risk attitude), then everyone who is capable (i.e. every moral agent) is allowed to do the exercise using one's own preferences.

Furthermore, each moral agent has his/her own welfare function W^a , but the welfare functions of different moral agents are not mutually well-calibrated. Different moral agents might use different scales or units for e.g. the levels of well-being. As estimates of well-being are not interpersonally comparable, a level of well-being equal to 100 according to moral agent a might mean a level equal to 10 according to moral agent b . As a result, the welfare function used by moral agent a might have much higher values than the welfare function used by moral agent b . How do we compare those welfare functions when they are not properly gauged?

These problems can be solved as follows. Consider the set G of all objective (i.e. quantifiable) distributable goods and burdens. These are goods and burdens in the sense that they can positively or negatively influence someone's well-being, they are distributable in the sense that we are able to distribute those goods between sentient beings (the goods are in our direct control, whereas someone's well-being is not in our direct control), and they are objective in the sense that they should be measurable (quantifiable). The latter property is important, because we want to avoid subjective estimates that differ between moral agents. Examples of such objective distributable goods are resources (economic wealth, income, energy, materials) and liberties (e.g. primary goods according to Rawls (1971), capabilities according to Nussbaum (2000)). The goods are subject to some constraints, such as maximum available resource levels and logically possible distributions of liberties.

Moral agents can now do the thought experiment as sincerely as possible, distributing the goods and burdens to maximize their own preferred welfare functions, respecting the constraints on the goods. Each of the momentaneous minds that compose person i gets a part of the distributed goods compatible with the lifetime well-being $x_i^a(g_a)$ where g_a is the distribution of goods (for example a person i gets a certain income at a certain time). Each moral agent a can now calculate the optimal distribution of goods g_a^{opt} that maximizes moral agent a 's welfare function at W_{max}^a . As G is a compact (bounded) set, the maximal welfare function is finite (it cannot grow to infinity).

We can now maximize the average of weighted welfare functions

$$\bar{W} = \frac{1}{N_a} \sum_{a=1}^{N_a} \frac{W^a}{W_{max}^a},$$

with N_a the number of moral agents. This averaging means that the subjective estimates of all moral agents count equally.⁴⁴ Using the relative welfare functions (W/W_{max}) implies that the optimal distribution of goods according to moral agent a is as valuable as the optimal distribution of goods according to moral agent b . This solves the problem of gauging the welfare functions of different moral agents. Even if moral agents might use different scales or units for e.g. the levels of well-being, taking the relative welfare function gives a good calibration for the welfare functions.

This is the democracy in impartial preferences of moral agents (a slightly different approach was proposed in Moreno-Ternero & Roemer, 2005). It turns a collection of subjective estimates into an objective, impartial solution.

In reality, this optimal impartial distribution of goods will be very difficult to determine, because each one of all the moral agents have to imagine being each one of all the current and future living sentient beings, experiencing each momentaneous well-being. It is like using the standard model of elementary particle physics to solve problems with many particles (e.g. to study the efficiency of a combustion engine). How do physicists deal with the complexity? Two considerations can make the principle easier to apply in daily life situations.

First, we can restrict the set of moral agent and sentient beings: in a lot of daily life situations, only a limited number of sentient beings are measurably (relevantly) affected

⁴⁴ There are three subtleties. 1) Do all moral agents (who exist and will exist here and everywhere) have a vote, or do only those moral agents who are able to influence the distribution of the respective goods have a vote? Ideally, I would say all moral agents should have an equal vote, but this is impossible in practice, so I go for the second option: only those moral agents who exist at the time of decision making and who can influence the decision, count. 2) What if the preferences (e.g. the level of risk aversion) of a moral agent changes over time? Either we take the average of the preferences of the moral agent over his/her life, or only the preferences at the moment of decision making. I would leave this choice up to the moral agents themselves. Most importantly, we have to be aware that the moral agents do not have cognitive biases at the moment of decision making. Effects of e.g. framing and priming or the influence of e.g. smells (triggering disgust) should be avoided. Experimental moral psychology can help us to determine all influences that generate such biases. 3) What if all except one of the moral agents do not feel an emotional problem with the resulting average distribution? Imagine that you would feel very unhappy when no-one except you has risk aversion behind the veil, i.e. when everyone chooses sum-utilitarianism whereas you really would want that we give some priority to the worst-off. For you, prioritarianism is very important. So imagine, that your well-being in a chosen sum-utilitarian ethic will be much lower than everyone's well-being in a chosen prioritarian ethic. If everyone has one vote, the result is close to a sum-utilitarian ethic, which has lower well-being due to your unhappiness. This loss of well-being should be taken into account, so a weighted democratic voting might be required. This is comparable to situations in physics, where a gravitational field created by heavy objects influences the path of another object, but the presence of the mass of this other object itself also influences the gravitational field to which it is subjected. This results in complex, non-linear interactions and effects. The (physical/moral) force (the gravitational field/the welfare function) is influenced by the objects (the masses/the decisions of moral agents).

by the policy of only a limited number of moral agents. These moral agents can hence apply the democratic approach, limiting the thought experiment of the veil of ignorance to only those sentient beings who are measurably affected. This is similar to what physicists do when they study interactions between a few number of particles (or a few number of planets and stars), assuming interactions with a background field of particles far away are negligible.

Second, if we know that moral agents do not have strongly divergent risk attitudes (that they can easily find a consensus on the issue of priority for the worst-off), and if the lifetime well-being estimates according to those moral agents are not so divergent when considering important, extreme cases (e.g. cases of extreme poverty or factory farms), we can easily derive rules of thumb to quickly decide the best estimate for the optimal distribution g^{opt} . Compare it with thermodynamics in physics: when a lot of particles interact, physicists use approximations (rules of thumb) to study e.g. the efficiency of combustion engines (I used the same analogy between branches of physics and approaches to ethics in section 3.5). These approximated rules are e.g. the laws of thermodynamics and fluid mechanics. In the end, they can be derived from the standard model of elementary particles, using a lot of statistical mechanics. But as engineers do not use the standard model of elementary particles to study combustion engines, ethicists and politicians do not have to use the complex welfare function all the time. They can derive equations or laws that are easier to apply in those daily life situations. And these new, derived moral rules can look completely different than the underlying QMM prioritarian principle, just as the laws of thermodynamics look completely different than the standard model in particle physics.

Of course, even when there was only one moral agent, s/he should appeal to approximate, derived moral rules, because it will be practically impossible to estimate past and predict future levels of lifetime well-being and to calculate the welfare function over a vast period of time. Science is necessary to tell whether the derived moral rules match the target principle of QMM prioritarianism.

4.7 Summary

In this section, I derived the quasi-maximin prioritarian principle as a model for consequentialist theories of justice. Along the road, we encountered multiple problems: how to define well-being, how to integrate well-being over a lifespan (when persons change), how to compare well-being between persons, how to distribute well-being between persons, how to deal with variable populations (e.g. future generations) and how to deal with uncertain outcomes (lotteries).

Quasi-maximin is a theory close to maximin, but with a small tendency towards sum-utilitarianism. The ‘quasi’ in QMM-theory is derived in two different ways: first from impartiality (the veil of ignorance) with a high but not maximal risk aversion (a high preference for security), and second from empathy (the equality principle) with a small but not zero preference for efficiency. QMM is a special form of prioritarianism that is compatible with the moral virtue of empathy and the moral intuitions of impartiality and efficiency.

Some implications of QMM-prioritarianism were discussed, and an elegant mathematical expression of the welfare function $W(h)$ of a world history h was derived using the veil of ignorance (with risk aversion for gains, risk neutrality for losses, and loss aversion), and using moral intuitions in population ethics. Quasi-maximin prioritarianism says that we have to choose the situation or world history which has the highest value of the welfare function.

Note that quasi-maximin, although a rational theory, already incorporates some moral intuitions in some subtle ways. In particular the level of risk aversion and the need for efficiency cannot be derived from purely rational reasoning. These non trivial (not zero or one) values for risk aversion and need for efficiency result into a prioritarian theory that best fits those intuitions (better than the extreme theories of sum-utilitarianism and maximin).

However, there are a lot of other intuitive judgments that are in conflict with and cannot be derived from the Rawlsian veil of ignorance. In the next sections, we will encounter some other moral intuitions that might overrule the QMM-principle. These moral intuitions are important in the ethics of care and the ethics of rights (Kantian deontological ethics).

Chapter 5 Partiality and tolerated choice equality

The consequentialist theory of prioritarian justice can be very demanding. Real impartiality might imply that we need to sacrifice many of our resources and much of our well-being in order to advance the worst-off individuals. Two replies can be given to this ‘demandingness objection’. First, we should require that governmental institutions and political laws are really impartial. Second, we note that for a moral agent some partiality can be tolerated under some conditions. Especially social or empathic beings have difficulties being perfectly impartial. We often have difficulties being impartial, because we have strong emotions towards our relatives, friends or co-living animals. This partiality might conflict with consequentialist theories like QMM-theory. Partiality can also be important for some moral patients. Imagine children growing up in a family of perfectly impartial parents. This will raise concerns about their emotional development and well-being.

According to an ethics of care (Noddings, 2002) we do not always have a duty to take the impartial point of view, because that would not respect interpersonal relationships. We have stronger empathy with people that we know well, and stronger personal involvement when we have a closer contact with someone. These emotions (of friendships) influence our decision making, which is not necessarily immoral.

To make the theory of justice less demanding, we can allow for some partiality. Partiality can trump the impartial theory of justice described in the previous section, under two conditions: 1) the violation of impartiality should not be too strong, and 2) the level of partiality should be universalized according to the Kantian categorical imperative: we should want to live in a world where everyone behaves with similar levels of partiality. These two conditions are related: if the level of partiality is too high, if the QMM-theory is too much violated, we would not want to live in a world where such levels of partiality are universalized. So partiality can weakly overrule the QMM-principle, and the principle of universalization (of section 1.2) is crucial. Partiality is

allowed to some degree as long as we are willing to respect similar levels of partiality of everyone else.

5.1 Tolerated choice equality

The inclusion of a tolerated partiality in ethics generates an equality principle which is different from the prioritarian equality principle of the previous section. For reasons to be discussed, I will call it ‘tolerated choice equality’.

Consider an example of the burning house dilemma. In animal rights discussions people sometimes give such a dilemma, whereby we have to choose between rescuing our child or a dog from a burning house (Gary Francione (2000) also referred to this dilemma in his book “Your child or the dog”). Of course the meat eater expects that also animal rights activists would save their own child, so they point at this kind of partiality to justify speciesism. The argument can easily be countered by changing the dilemma a bit: choose between your child and a child with another skin color. Of course, people are not necessarily racist when they prefer to save their own child. I will explain why not.

There is indeed an *emotional inequality* between children. But suppose you were at the burning house, and you chose to save the other child instead of mine. If I’d tell you that my child has a higher moral status and a stronger right to live, due to its skin color, I would be a racist. But if I tolerate your choice to save the other child instead of my child, I would not be racist. The reason is that I consider you and me to be morally equal, and the children in the house inherit this kind of equality. I tolerate your partiality, and therefore the children have inherited a tolerated choice equality which is not in contradiction with the emotional inequality that I feel. This new principle of equality can be formulated as follows:

The tolerated partiality principle. You are allowed to be partial as long as you tolerate similar levels of partiality for everyone else, and if the partiality is not based on false beliefs or prejudices.

More specifically: when helping others, you are allowed to give (to some level) priority to those with whom you feel a personal or emotional concern or involvement, on the condition that you should tolerate the choice of other caregivers to give priority to whom they prefer (their loved ones). So you should tolerate the choice of other helpers.

Tolerated choice equality. If 1) you want to help a person X (for whom you feel an emotional concern) and another helper wants to help person Y, if 2) you consider the other helper as being equal to you, and if 3) you tolerate the choice of the other helper to help Y, then persons X and Y inherit a tolerated choice equality.

5.2 To whom applies the tolerated choice equality?

In this section I try to answer the question who we need to take into account for this new principle of equality? When we look at consequentialist principles such as QMM-prioritarianism, it was self-evident that all sentient beings should be taken into account, because well-being is what matters from behind a veil of ignorance.

Suppose in the burning house dilemma I had to choose between saving a child, a dog or a car. In principle, as with the veil of ignorance, all entities in the universe should be taken into account, including cars. But if I saved the car, it could not be tolerated, because the QMM-principle will be violated far too much. Therefore, the tolerated choice principle should only be applicable to sentient beings. It ‘inherits’ this criterion of sentience from the QMM-principle, and in the QMM-principle it was derived from the veil of ignorance. Cars are not sentient, so we cannot influence their well-being, no matter what we do. Another reason why the tolerated choice principle is applicable to all and only to sentient beings is that the principle stems from feelings of personal connection and empathy with others. Of course feeling empathy only makes sense towards sentient beings. We cannot feel empathy with a car.

It is important that this partial aid or care is motivated by feelings of empathy and concern. Tolerated choice equality therefore meets some criticism by proponents of an ethics of care, who claim that impartiality is too ‘cold’.

Going back to the dilemma between saving a child or a dog, a true antispeciesist should have to tolerate the choice of someone who saves the dog. Note that a lot of people give more food and medicines to their pet dog than to starving children, but those choices are already tolerated to some degree. According to the QMM-principle, we would have to calculate the well-being of all individuals involved; whereby we have to take into account the life expectancies and potential levels of well-being (some sentient beings have a richer and more complex emotional life than others). So it might be argued (from behind the veil of ignorance) that saving a mentally healthy human child would better correspond with the QMM-principle than saving a dog. Nevertheless, the principle of tolerated choice says that some slight violations of the QMM-principle should be tolerated. Otherwise we end up with a too demanding impartial view that is in contradiction with some of our strongest moral intuitions. You might save a mentally disabled child instead of a mentally healthy child, even if the disabled child will have an opportunity for well-being as high as a dog, i.e. lower than the healthy child.

Let’s give another dilemma to clarify some points. A trolley is moving at great speed. On the main track you see someone you hold dear (e.g. your child or your partner). You are standing next to a switch. You can turn the switch to save that beloved person, but on the side track there is another person whom you do not know. You have the choice between doing and allowing harm. If you do nothing, you allow your beloved person to

be harmed. If you turn the switch, your action causes harm to someone. According to our theory, there is no morally relevant distinction between doing versus allowing harm in this set-up. You are allowed to turn the switch to save that beloved person.

What if on the side track there are two persons? Are you allowed to be partial to such a degree that you cause more harm by your action? If you turn the switch, I can understand your choice, so I might tolerate it a little bit. But only a little bit, because your action definitely violates the QMM-principle. My intuition says that we should not cause more harm by saving someone we hold dear. But there is another subtle slightly similar situation: imagine there are three tracks. If you do nothing, the trolley will take the main track and kill five unknown people. You can turn the switch, so the trolley takes the second track and will kill someone you hold dear. The third alternative is turning the switch further, sending the trolley to the third track where it will kill two unknown people. In this case, you are allowed to turn the switch to save the five people on the main track. But my intuition says that you are allowed to send the trolley to the third track. I.e. you do not have to choose the second option that saves most lives and best respects the QMM-principle.

In section 6.1 I will elaborate more on these kinds of dilemmas. I will put forward another principle that backs up the above intuition that we are allowed to send the trolley to the third instead of the second track, killing two people instead of one. And I will demonstrate that there are situations where we are not only allowed to tolerate the partial choice of a helper, but that we *should* tolerate it in order to respect the helper.

What is relevant in this distinction between doing and allowing is the intention. Suppose on the main track there are five people whom you don't know, and on the side track there is one person whom you really hate. You always wanted to kill that person, so now you see your chance to do so by turning the switch. The trolley will take the side track and kill that hated person. The action itself is allowed because it is in agreement with the QMM-principle. But the intention is wrong. So you would not be punished for turning the switch, but you are a risk to society by having malicious intentions. We can't trust you anymore to respect the QMM-theory in the future. Imprisonment might be necessary, not because of a punishment, but because of a protection. You have a duty to change your malicious intentions.

I have given examples to indicate that some of our moral intuitions say that the tolerated partiality principle weakly overrules the QMM-theory of prioritarian equality. This adaptation of the theory is not inconsistent, it is better in line with our intuitions, and therefore I consider it a better theory. In Chapter 6 we will encounter moral intuitions that generate a principle that more strongly overrules the QMM-theory as well as the tolerated partiality principle. Before we move to that section, I briefly discuss a possible unification of tolerated choice equality and equality of opportunity.

5.3 Tolerated choice equality and equality of opportunity

The principle of equality of opportunity (Arneson, 2008) says that all people who are equally qualified should have equal prospects for benefiting positions such as jobs. Jobs should go to the most qualified (e.g. most talented and motivated) persons and not to persons for arbitrary, irrelevant reasons such as sex or race.

Tolerated choice equality says that we are allowed to be partial to some degree. If a heterosexual man prefers to marry a woman instead of a man, one might say that this heterosexual man violates the equality of opportunity between women and (homosexual) men. So the heterosexual man is partial towards women, and we tolerate such partiality. The heterosexual man is not sexist if he tolerates the choice of homosexual men to marry homosexual men. If he says that no man is allowed to marry a homosexual man, then it would be sexist.

Now consider a white employer who chooses a white job applicant instead of a black person. This choice might violate equality of opportunity, and it might be a racist kind of discrimination. Is the employer willing to tolerate similar degrees of partiality of everyone else? The employer might say that s/he would tolerate the choice of another employer to hire a black person. But this does not yet guarantee tolerated choice equality, because in a competitive market the employer might look at other employers as competitors, and hence as being unequal in some sense (see the condition in the above formulated principle of tolerated choice equality: you should consider the other helper as being equal to you). The employer might have prejudices towards black people, thinking that black people are not good workers. So the employer might be glad to know that other employers hire such bad workers.

The above implies that in competitive environments, tolerated choice equality cannot always be derived: if employers are competitive, an employer cannot justify a partial choice (a partiality towards some job applicants) by tolerating the choices of the other employers. For the heterosexual man, there was no competition with homosexual men, so tolerating the choices of those homosexual men generates tolerated choice equality between men and women. The same goes for a man who is more sexually attracted to white women than to black women. He prefers to marry a white woman, but if he tolerates the choices of other white men to marry black women, especially if he considers those other white men as his equal (for example if his brother wants to marry a black woman), he would not be racist.

It is allowed to violate equality of opportunity; as long as there is a tolerated partiality (tolerated choice equality). We can further explore the difference between a racist employer who prefers to hire a white job applicant, and a non-racist man who prefers to marry a white woman. The motivation becomes important: what drives the employer to prefer white people? If the employer is afraid of black people, s/he is not

necessarily racist. But if s/he has prejudices towards black people, it becomes discrimination. The employer can overcome those prejudices. The man who prefers white women can also overcome his preference for white women. But this preference is much more difficult to overcome than the prejudices. Changing opinions is easier than changing taste preferences. The tolerated partiality principle refers to personal or emotional concerns, and such concerns are not easy to overcome.

The above can be summarized in the following combination of tolerated choice equality and equality of opportunity.

Tolerated choice equality of opportunity. Suppose persons A and B offer two similar positions¹ and persons X and Y compete for the position offered by A. A prefers to give the position to X. Then A respects the tolerated choice equality of opportunity if the following conditions are satisfied: 1) A considers B as being equal (excluding competition between A and B), 2) A would tolerate the choice of B to give B's position to Y, and 3) A's preference for X is not based on false beliefs (prejudices) but on taste preferences that are not easy to overcome.

¹ Here, the offered positions can be interpreted as jobs, permissions to marry someone,...

Chapter 6 Basic right equality¹

6.1 Moral dilemmas and strong moral intuitions

In a previous chapter (3.3) we encountered some trolley dilemmas. Those dilemmas were an interesting tool to demonstrate the importance of deontological judgments. In the first dilemma, referred to as the switch dilemma, a trolley is going to kill five people on the main track. However, you can hit a switch so that the trolley takes a side track, to save those five people. Unfortunately, on this side track there is one person. The structure of the dilemma is: doing nothing results in the death of five people, acting (pulling the switch) results in the death of one person. Our theory of prioritarian justice states that one person dead is better than five people dead, so we should turn the switch.

However, in another dilemma, referred to as the bridge dilemma (similar to situation C in the previous chapter 3.3), again a trolley is about to hit five people. You can push a heavy man from a bridge in front of the trolley in order to block the trolley. The heavy person will die, but the five people on the track will be saved. A lot of people have the intuition that pushing the heavy man is not allowed (Hauser et al., 2008).

In a third dilemma, called the hospital dilemma, five patients in the hospital need new organs in order to survive. However, no organs are available anymore. Is it allowed to kill a visitor (against his will) and use his kidneys, liver, heart and spleen for transplantation to save the five people? Here as well, most people are deontologists: they are very reluctant to allow such actions, even if they could save more lives.

What are the distinctions between those dilemmas's that can explain the different judgments? As Greene (2001; 2004) pointed out, there is a difference between up-close-and-personal situations (pushing the heavy man), and more distanced/detached

¹ Sections 6.2, 6.3 and 6.6 are based on *Can Deontological Principles be Unified? Reflections on the Mere Means Principle*, submitted to Philosophical Studies

situations (pulling a switch). This results in an emotional inequality that we can tolerate (see previous section about the tolerated choice principle). But there is more. What if we did not have to push the heavy man, but simply push a button that will topple the heavy man from the bridge? My intuition says that even then, action is not permitted.

6.1.1 A first approach: uncertainty aversion

Let us look at this bridge trolley dilemma from behind the veil of ignorance. Suppose you don't know who you will be: you can be any of the six persons involved (the heavy man on the bridge and the five people on the main track). You can now decide between two possible worlds. In the first, the heavy man will not be pushed, in the second he will. Which world would you prefer? If you are really sure that the plan to block the trolley by the heavy man will work, you would rationally speaking rather be in the second world, because your chances of survival are five times higher. Only if you were the heavy man, you would die. In the first world, you would die if you are one of the five people on the track. But now suppose, as in real life situations, you are actually not sure that the plan of blocking the trolley will work. Perhaps the trolley is too fast and the heavy man not heavy enough to stop it? Then one or more people on the track would die as well.

We already mentioned that from behind the veil of ignorance, we might have risk aversion that results in a quasi-maximin strategy. But risk aversion implies that we know the probabilities of survival. In this case however, we don't even know the probability of the plan to work. We don't know the chance on survival. It is a situation, not of risk, but of uncertainty (or ambiguity).

In order to understand the effects of uncertainty, let's consider the example of the Ellsberg paradox (Ellsberg, 1961). An urn contains 60 balls. You know that there are six different colors of balls, and that there are ten green balls. That's the only information you have. You can now choose between two games of chance. In the first game you win when you draw a green ball. In the second game you win if you draw a blue ball. People who have strong uncertainty aversion prefer the first game, because then they at least know their probability to win ($1/6$). In the second game, they only know that their probability is somewhere between 0 (if there are no blue balls) and $5/6$ (if there are only green and blue balls). So people can not only have risk aversion; they can have uncertainty aversion as well.

Looking back at the trolley dilemma, we also have a choice between two games of chance. In the first game (the world where the heavy man is not pushed), you have a probability to survive (to win) equal to $1/6$. In the second game, you don't know your probability of winning. You have uncertainty about probabilities, and if you have an aversion for such uncertainties, you'd prefer the first game.

We see that the veil of ignorance already comes pretty close to a lot of our moral intuitions. First, it values impartiality and well-being, as in a consequentialist theory. Second, having a high but not maximum risk aversion, we arrive at a prioritarian justice, which is coherent with our empathy and small but not zero need for efficiency. Third, having uncertainty aversion, we arrive at moral judgments that correspond with some deontological moral intuitions. In a later chapter on the predation problem (see Part III), we encounter another implication of uncertainty aversion from behind the veil of ignorance, which might be able to explain our tolerance towards predation.

We might ask ourselves the question how much risk aversion and uncertainty aversion we should (or are allowed to) have. The veil of ignorance does not provide an answer. But we could introduce a second veil. Behind the first veil, you are a moral agent who does not know which being in the real world s/he will be. Behind the second veil, you do not know what kind of moral agent sitting behind the first veil you will be. So imagine that you are behind a second veil: you know that you will soon be a rational (moral) being that is about to perform a thought experiment of the veil of ignorance. But at this moment you don't know how much risk and uncertainty aversion you will have. You do know that most rational beings (moral agents) have some risk and uncertainty aversions (because based on psychological studies, most humans have these aversions). So likely you will also get a high level of risk and uncertainty aversion when you are behind the first veil.

From behind this first veil of ignorance, having uncertainty aversion, you might prefer a situation where the QMM-principle might be violated if the alternative would be a situation of uncertainty. But not all has been said yet. Our moral intuitions say that we are not allowed to push the heavy man or sacrifice a visitor in the hospital for transplantation, even if we can be very sure that the plan of saving five other people (the people on the track or the patients in the hospital) will work. So let's look for a universalized ethical principle that clearly expresses these moral intuitions. I will first criticize some tentative accounts encountered in the literature.

6.1.2 Tentative ethical principles

In a previous chapter we encountered some possible explanations for the differences in moral judgments in the trolley dilemmas. Some people (Boorse, 1984; Harris, 2000; Postow, 1989; Waldmann and Dieterich, 2007) proposed that there is a morally relevant distinction between sending a trolley to the victim (which is done in the switch dilemma) and sending a victim to the trolley (which is done in the bridge dilemma by pushing the heavy man). But we demonstrated that this was a moral illusion.

Also in relation to the hospital dilemma, people make a distinction between death by an existing threat (e.g. an organ disease) and introducing a new threat (killing a visitor

with a knife). But such differences appear to be artificial constructions. Consider the following dilemma. Five persons are on a moving platform on the rails. If you do nothing, the trolley will crush the platform and kill those people. But you can move the platform away from the rails in order to save the five. But this move will push another person (who is next to the platform) to an electric cable. This one person will consequently die by electrocution. I believe that action is allowed. But here we see that first a new threat is introduced (the electric cable), and second the victim is pushed towards this threat (the cable is not moved towards the victim).

Some people (e.g. Kamm, 1989) say that there are morally relevant differences between the causal chains in the switch and the bridge dilemmas. In the bridge, the action first results into threatening and harming the heavy man, and after that the five people are saved. In the switch dilemma, the action simultaneously saves the five on the main track and threatens the one on the side track. The harm done to the one person on the side track occurs later in the causal chain, compared to the harm done to the heavy man. So the structure of the causal chains is different. The problem with this approach is that we can invent dilemmas such as situation B encountered in a previous chapter (3.3), where it becomes complicated to see what the morally relevant aspects of the causal chain are. And neither does it seem really relevant when harm is done in a causal chain. A ‘causal myopia’ might also be a moral illusion just like we demonstrated that an intervention myopia was an illusion.

Some people (Reibetanz, 1998; McIntyre, 2001; Fischer & Ravizza, 1992; Shaw, 2006) refer to intentions and the Doctrine of Double Effect to justify the differences between the switch and the bridge dilemmas. The doctrine says that there is a moral difference between the intentional harm as a means versus the foreseen harm as side-effect (Quinn, 1989b). One critique is that this doctrine refers to the moral difference between an intentional harm as a means and a foreseen harm as a side-effect. However, the interpretation of this difference becomes very difficult in e.g. the loop trolley dilemma (discussed in section 3.3). So the interpretation of this doctrine, and its application to trolley dilemmas, is not clear. In the appendix, this point is discussed a bit further. A second critique is that actions can be permissible even when agents have bad intentions. For example when person A hates person B who is on the side track, and person A turns a switch that sends a runaway trolley to person B in order to kill him, this act is permissible if turning the switch implies saving the people on the main track. Person A’s intentions and moral character are bad, but my intuition says that the act is good.²

² This intuition is consistent with the universalization criterion that focuses on rules instead of acts (“You may *follow the rule* that everyone may follow in similar situations.”). The criterion says you may do an action if you can find a rule that justifies the action, if the rule is compatible with the ethical system and if you can want to see this rule universalized. Suppose I kill a person on the side track. I am allowed to do this action if I

All the above tentative explanations and principles could be applied to animal ethics, even if they are moral illusions. But I prefer a clear criterion that distinguishes between the dilemmas, a criterion that can be translated in something morally relevant, such as a right. If we could refer to such a right, then we might arrive at a new principle of equality, where everyone has an equal claim to this right. The right not to be killed will not do, as in both trolley dilemmas an innocent person will be killed when the agent acts (when he pulls the switch or pushes the heavy man). As we have seen in a previous chapter, another special right is able to do the job.

6.2 The basic right and the mere means principle

The right that solves the problem of the difference between the switch and the bridge dilemmas is the deontological or basic right not to be used as merely a means to someone else's ends. We can see that the heavy man will be used as a trolley blocker (human shield), and that the visitor in the hospital will be used as an organ donor. But the person on the side track will not be used as merely a means to save the others. One does not need that person in order to pull the switch and save the five.

This basic right does not follow from the veil of ignorance, although it can be compatible with it in situations of uncertainty mentioned above. Neither does the basic right follow from a feeling of empathy. The basic right has different origins, and is highly coherent with many moral concepts and moral intuitions.

1) The basic right stems from a feeling underlying *respect*, which is considered a moral virtue to be developed. Treating someone as merely a means is not respectful.

2) The basic right is related to the notion of *intrinsic value*, which is to be distinguished from instrumental value. We give something intrinsic value when that thing is important (valuable) beyond its use value.

3) The basic right is related to the notion of *dignity* and resembles a version of the Kantian categorical imperative: "Act in such a way that you treat humanity, whether in

can find a justifying universalized rule. A rule that will not work, is: "Turn switches in order to kill people you hate." If I only have such a rule, I am not allowed to act. But I can find another rule, such as "Turn switches in order to satisfy the prioritarian theory, as long as no other principles of the ethical system are violated." I can do the same act (turning the switch) by following this rule. This universalized rule can justify turning the switch, even if in reality I happen to have a bad intention. In other words: a (bad) intention is not important in judgments of permissibility of an action if there is a justifying rule for that action that does not refer to the (bad) intention.

your own person or in the person of any other, never merely as a means to an end, but always at the same time as an end.” (Kant, 1785) According to Kant, humans have dignity. But as I am going to demonstrate, there are good reasons why not only humans but also other sentient beings have dignity and should get a basic right.

4) The principle of the basic right is a universalized ethical principle that is consistent with our *moral intuitions in a lot of moral dilemmas*. I will list ten of them.

a) The trolley dilemma: we are not allowed to push a heavy man from a bridge, to fall in front of a runaway trolley, in order to block the trolley that is about to kill five people on its track. We should not use a person as human shield (Thomson, 1985).

b) Organ transplantation: we are not allowed to sacrifice a person against his will, using his organs in order to save five patients in the hospital who will die without new organs.

c) Human cannibalism: survivors in a lifeboat should not sacrifice and eat another person in order for them to stay alive. We should not use a person as food.

d) Involuntary experimentation: we are not allowed to perform experiments on a person in order to find a therapy that will save many people. We should not use a person as laboratory equipment.

e) Terror bombing: we are not allowed to kill a few innocent civilians in order to demoralize the enemy, win the war and save more lives.

f) Torture interrogation: we are not allowed to torture a person in order to gain information about a bomb that will kill many people (especially when the person is an innocent eyewitness who discovered the location of the bomb but is threatened by terrorists not to reveal the location, or when the person is the innocent child of the terrorist and the terrorist will only reveal the location when his child is tortured).

g) Blackmail murder: we are not allowed to kill an important person if a terrorist threatens to kill five hostages instead. We should not use a person as ransom.³

h) Nude photography: we are not allowed to take and sell nude pictures of a person against his/her will in order to satisfy thousands of porn consumers.

i) Gang rape: a woman should not be forced to have sex with frustrated men. We should not use someone as sex toy, even when the increase in total well-being of the rapists would more than compensate for the loss of well-being of the victim.

³ This example is similar to the thought experiment proposed by Williams (Smart & Williams, 1973, p.97-100): should Jim kill one Indian if refusing to kill this Indian implies that armed men will kill twenty Indians instead?

j) Human zoos: we should not lock up a strange looking person in a circus or zoo for the entertainment of many other people.

k) Scapegoat: we should not prosecute an innocent individual, even if such a prosecution would stop a riot that will kill many people.⁴

l) The sadistic conclusion: we should not allow the birth of someone who will have a life not worth living, even if the (weighted) average well-being were negative (i.e. even if the welfare function would increase by introducing a life that is not worth living, see section 4.6.6.4).

Many more situations can be given, such as trafficking (buying and selling humans), slavery or gladiator fights. All these practices have something special in common: a person is used as a means (as a human shield, trolley blocker, experimental object, deterrent, information source, ransom, sex object, toy, painting, scapegoat,...) against his or her will. Many people have the intuition that the above practices are impermissible, even if the overall consequences in terms of lives and well-being would be better. The more examples we can give of situations where an action is not allowed when the victim is used as merely a means, the less likely it is that our intuitive moral judgment of impermissibility is a moral illusion. The coherence between those dilemmas gives credibility to our moral intuitions and our corresponding universalized ethical principle.

5) The basic right is also related to a restricted kind of *propertarian libertarianism*. This kind of libertarianism is based on *bodily autonomy*, as a special property right over one's own body. In contrast with most libertarian theories, this restricted version allows some space for distributive justice and avoids the conclusion that taxation is comparable to slavery. This will be discussed in the next section (6.3).

6) The basic right is a formulation of the *mere means principle*. It can be interpreted as a specification of the *doctrine of double effect* (McIntyre, 2011). This doctrine of double effect (DDE) refers to an action that has a good and a bad effect. One of its crucial conditions states that such a double effect action is not allowed if the bad effect is intended as a means to the good effect (or if it is intended as an end in itself). The problem with the DDE is that in some moral dilemmas it is not always clear what it means to intend (rather than foresee) a harm as a means to an end. The basic right (mere means) principle avoids the intention-foreseeing distinction and instead specifies the means-end distinction. As we will see, this is done by a more algorithmic procedure:

⁴ This example is similar to the dilemma of 'framing the innocent man' (McCloskey, 1965): a race riot (angry white people retaliating and killing black people) can be stopped by quickly arresting an innocent black person, bearing false witness and punishing him in order to quiet down the situation.

the counterfactual question whether the end could be achieved if the victim was not present.

7) As I will discuss below (section 6.6), we can extend the mere means principle in a way that it is not only immoral to *use* but also to *consider* someone as merely a means. This extension can explain other deontological intuitions such as the difference between *doing versus allowing* (Howard-Snyder, 2011), the permissibility of *partiality in imperfect duties of beneficence* (Beauchamp & Childress, 2011), and the *asymmetry of procreational duties* (Narveson, 1967; Mulgan, 2006, McMahan, 2009). It is as if this extended mere means principle unifies a lot of deontological principles and intuitions.

In section 3.4, I already demonstrated that the deontological basic right is not necessarily a moral illusion. If we now look at the above points, we see a strong coherence of the basic right principle with moral virtues (respect), moral intuitions (in at least ten different dilemmas, as well as in situations of imperfect duties and procreational duties), moral concepts (intrinsic value, bodily autonomy, libertarian property) and deontological principles (double effect, doing versus allowing). This strong coherence indicates that the intuitions underlying the basic right principle are not moral illusions. The basic right principle is not arbitrary, artificial or farfetched. The principle can be clarified, as I will do in the next section, so we avoid fuzziness as well. And looking at the totally different situations where it applies (the abovementioned ten dilemmas), we see that it is not really context dependent. Another argument to see why it is not context dependent is that the basic right is something that individuals always have, independent from the situation. (Compare it with the fact that in the Müller-Lyer illusion a length is something that line segments intrinsically have, independent from their environment). This is something different than the tentative solutions presented in the previous section (and chapter 3.3), such as the ‘protophysical’ explanations.

Of course, the above is not solid proof, the basic right might still stem from moral illusions. Note that not everyone has the intuitions underlying this basic right. But these intuitions seem to be culturally independent. It has more to do with different brain (mal)functionings (Greene et al., 2001). Most people (most moral agents) have these intuitions, and they are likely inborn. People who lack those intuitions might still be able to derive a consistent ethic, which will more resemble a utilitarian/consequentialist ethic. For those utilitarians, animal equality has to be applied as well (e.g. Singer, 1975). Here, however, I follow and respect those ‘basic right’-intuitions that the majority of moral agents appear to have.

Looking at the formulation of the basic right – use someone as merely a means for someone else’s ends – we have to answer three questions:

- 1) What do we mean by use as ‘merely a means’?
- 2) What do we mean with ‘ends’?

3) Who is the ‘someone’? I.e. who gets the basic right?

The first question will be answered in the next section. Questions 2 and 3 are related and will be dealt with in the subsequent section (6.4).

6.3 When is the basic right violated?

The mere means principle finds its roots in Kant’s categorical imperative (Kant, 1785): never treat a person merely as a means to an end, but always at the same time as an end. Yet, this landmark principle in deontological ethics lives on being (re)interpreted and discussed until today (for recent work, see e.g. Scanlon, 2008, ch. 3; Parfit, 2011, ch. 9; Kerstein, 2009).

When do we use someone as merely a means? Slavery, human trafficking, rape, cannibalism, involuntary organ donations, involuntary human experiments and pushing a heavy man from the bridge in order to stop a trolley are all examples of basic right violations of humans. What have these situations in common? And how to distinguish these examples from actions that do not violate the basic right? E.g. using a baker to get some bread, using an employee, sending your children to school against their will, imprisoning a criminal or killing a person on a side track in order to save five people on the main track. These actions are not immoral, and therefore should not be classified as basic right violations.

6.3.1 Two words, two conditions

The ‘mere means’ principle that generates the basic right, contains two words. Hence, two conditions need to be satisfied. The first condition says that an agent (the user) forces (in the broadest sense) the victim to do or undergo something against the will or interests of the victim (for example the victim does not want the treatment). This is the ‘mere’ part. The second condition says that the presence of the victim is required in order to reach an end of the user or someone else.⁵ This is the ‘means’ part.⁶

⁵ If the victim undergoes something against his/her will for an end of the *victim*, s/he is not used as merely a means for someone else’s ends, and the mere means principle is not violated. For example keeping a scared patient in a hospital for his/her own sake is not impermissible. This is particularly true in wildlife rescue centers, where injured or ill animals are kept against their will in order to help them. Paternalism might

The first condition, the ‘mere’ part of the mere means principle, is something a (rule⁷) utilitarian or consequentialist can agree with: doing or undergoing something against your will generally lowers your well-being and can be considered as a harm. It is the second condition that gives the mere means principle its deontological flavor. Looking at the above examples, we see that the presence of the victim is required in order for the plan to work. If the innocent civilian was not present in the terror bombing situation, the enemy would not become demoralized. If the important person was not present in the blackmail situation, you could not kill him and use his death as a ransom to free the hostages. If the fat man was not on the bridge, you could not push him and use him as a shield to block the trolley. On the other hand, consider the switch trolley dilemma: most people claim that when a runaway trolley is about to kill five people on the main track, we are allowed to turn a switch in order to redirect the trolley onto a sidetrack, where the trolley will kill one person. Although killing this one person likely goes against her will, she is not used as a means, because her presence is not required in order to turn the switch and save the five people on the main track.

The same goes for other situations where a victim “could not possibly consent” (to use Korsgaard’s expression (Korsgaard, 1996, p.138)). A rule utilitarian can agree with her: if the rule (the maxim) of our action precludes the possibility of the victim’s consent (the victim could not rationally will to be treated that way), we are not allowed to treat the victim that way. Korsgaard gives the examples of deception and coercion. A rule utilitarian might prefer to stick to the rule “do not lie”, because such a rule generally promotes well-being. But in contrast to Korsgaard’s view, I think the mere means principle is only violated when the liar wants the presence of the deceived person in order to reach someone else’s end.

Another example is imprisonment: violating someone’s liberty without consent. Imprisoning a murderer does not violate his basic right, because the presence of the criminal was not necessary in order to reach the end (a safe society). On the contrary, his absence was preferred. Imprisoning him might be the best strategy to reach the

sometimes be immoral, but not in the wildlife rescue center, because it does not violate the mere means principle and it promotes well-being.

⁶ These two conditions come close to the two conditions in Bognar & Kerstein (2010 p.15). 1) “A person treats another person *merely* as a means if it is reasonable for her to believe that something she has done or is doing to the other person renders that person unable to consent to her treating him as a means to her aim.” 2) “A person treats another person as a means if she intentionally does something to the other’s body or mind in order to realize one of her ends and she intends the other’s body or mind to contribute to her end’s realization.” This condition refers to the required presence of the body.

⁷ A rule utilitarian looks for those rules that, when they would be consistently respected in all similar situations, would in general generate most well-being. A rule utilitarian prefers to stick to the rule even if in a particular situation a violation of the rule would promote well-being.

quasi-maximin principle, even if we deprive his liberty. But using this murderer for forced labor violates his basic right.

The next two sections explain the two conditions in more detail.

Condition 1: the agent's behavior violates the interests of the victim.

When the victim has a will, this condition says that the victim does not want the agent's behavior⁸. For autonomous beings, this condition refers to autonomy and consent, but there is a whole spectrum of possible interpretations of autonomy. At the two extremes, there are stronger (narrower) and weaker (broader) interpretations of the 'mere' part in the mere means principle.

What is consent?

The strongest interpretation says that the victim is used as *merely* a means when s/he is not able to give rational, informed consent. I refer to Beauchamp and Childress (2001, ch.4) for a discussion on consent, but it is clear that according to this interpretation, the mere means principle is only applicable to rational beings: beings who are able to understand relevant information and give free consent. This is the traditional interpretation of Kant (1785), Korsgaard (1996) and many others.

The weakest interpretation says that the victim has to do or undergo something that s/he does not want.⁹ In this interpretation, not wanting something means: having negative emotions about it (or having a negative attitude towards it). Positive and negative emotions indicate that a being has subjective preferences or interests. For example, when pain generates a loss of well-being, it indicates that the individual wants to avoid bodily injury. Fear indicates a need for safety, and similar needs or interests lie behind other emotions. The mere means principle now becomes applicable to all sentient beings, i.e. beings who have developed (and not yet permanently lost) the capacity to experience positive and negative feelings that indicate the satisfaction of preferences. The advantage of this weaker (broader) interpretation is that the mere means principle is also applicable to mentally disabled (a-rational) humans. This corresponds with the intuition of many people.

⁸ Of course the agent's behavior has to be related to the use of the victim. If I buy bread from my neighbor, I use my neighbor as a baker. If I annoy my neighbor when I put my music loud, I do not use my neighbor as merely a means, because the loud music is not related to buying bread. However, if my intention is to annoy my neighbor with loud music in order to coerce him to bake bread for me, it will become a use as merely a means.

⁹ We should understand this in a broad sense, which includes not wanting deception. A deceived victim might not actually experience negative emotions, but if this victim does not want to be deceived that way, it counts as a violation of consent.

In the next section on who gets the basic right, I discuss the possibility of an even weaker interpretation, leaving the notion of consent behind, and focusing exclusively on interests. This is a very broad interpretation, because now non-sentient beings with interests (e.g. living beings such as plants) can be victimized.

Why consent?

According to a consequentialist welfare ethic, consent is important and counts in the utility calculus. But the consequentialist does not see a difference between the consent of someone whose presence is required, versus the consent of someone else. So the difficult question becomes: Why is the consent of the person whose presence is required so much more important than the consent of the person whose presence is not required? The deontologist has difficulty answering this question. Of course, s/he can refer to the coherence of his/her moral intuitions in the situations given in a previous section. Alternatively, s/he can give a rationale such as: if presence becomes important, autonomy dictates that consent becomes especially important. But as the next example demonstrates: it is not the general (lack of) consent of the victim that is important. Only the consent about the presence of (the behavior of) the user has a special status.

An example from economics: the poor baker

The condition that the victim does not consent with the agent's behavior is important. Consider a poor person who decides to work in a bakery. He hates getting up early in the morning to bake some bread, but his poverty gave him no choice except bake or die. If I buy his bread, I am using him: his body is necessary to make the bread. But although he hates baking bread, the poor baker does not have a negative attitude towards my behavior. In other words: my behavior did not cause his poverty. If my behavior was not present, the baker would still be poor. Therefore, I am not using him as merely a means. On the other hand, if I threaten or force someone to work in a bakery, it becomes slavery and I am causally responsible for his bad situation. The agent causes a violation of the rights of the victim, if the presence of the agent is a necessary condition of the harm.

Timeframe of the agent's behavior

A tricky question concerns the boundaries of someone's behavior. Consider a slave owner who claims that his slaves are better-off as slaves than they would be in the wild, because as indigenous people in the wild they would face predators, diseases, drought, hunger and other nasty things. The owner protects the slaves and gives them food. So it might be true that a slave would prefer a life as a slave over a life in the wild. Hence the slave prefers the total behavior of the owner over the complete absence of the owner, if

absence means a miserable life in the wild. According to a broad interpretation of the behavior, this slave is not used as merely a means.

But based on my moral intuitions, I prefer a narrow interpretation that focuses at a particular behavior at a particular time. According to this interpretation, the slave is used as merely a means as soon as the owner does a particular thing that the slave does not want, for example whipping the slave (even if the slave prefers the total life of a slave with whipping over the alternative life in the wild). The same goes for the practice of breeding slaves: even if a slave would prefer the life of a slave over the absence of a life (the slave would not have been born if the owner did not breed slaves), it does not mean that the slave is not used as merely a means when s/he is whipped.

The latter resembles a situation of livestock farming: what if the life of a cow raised in a humane livestock farm is better than no life at all and better than a life in the wild, but the cow is still slaughtered for meat? For the cow, the procedure of breeding, raising and slaughtering, considered as a whole, might be preferable to not being born at all or being born in the wild (and e.g. being eaten by a predator at an early age). But we should not look at the procedure as a whole: it is the act of slaughtering itself that violates the mere means principle. Slaughtering (for meat) is a single act that the cow does not want and where presence of the body of the cow is necessary. According to the non-consequentialist (non-welfarist) mere means principle it is better that a cow is not born at all than that a cow with a life worth living is used as merely a means when she is slaughtered. This is the same logic as with the whipping of human slaves who have a life worth living. We do not have a duty to breed and raise happy cows (see the section on the asymmetry of procreational duties below), but once we cause the birth of a happy cow, we should not violate its basic right. We should not slaughter and eat the cow.¹⁰

It might be the case that other moral agents have another intuition than I have, that they want to take a broad interpretation of the mere means principle, judging the morality of a use in terms of the total behavior instead of a particular behavior at a particular time. Those moral agents might conclude that some kinds of slavery and meat consumption are permissible: when the lives of the slaves and animals are worth living and when the alternative would be that those slaves and animals were not born. To deal with this difference in moral intuitions, a democratic decision procedure amongst all moral agents (everyone who is capable of understanding this moral problem) might be a way out.

¹⁰ McMahan (2008) also discussed this issue of humane farming, arguing that we should not kill a happy cow, even if the life of the cow being raised and killed might be better than no life at all. This 'Logic of the Larder' (purchasing animal products is good because it can increase the number of animals whose lives are worth living) was also criticized by Matheny & Chan (2005). One of their claims is that animal farms prevent positive lives of wild animals.

Condition 2: the agent wants the presence of the victim's body

The second condition of the mere means principle can also refer to a mental state of the agent, namely what the agent wants. This subjective mental state is connected to an objective fact: the presence of the victim that is causally required for the end of the agent. This can be tested by a counterfactual thought experiment: does the agent's plan still work if the victim was not present? If not, then the victim is used as a means.

But what exactly should be present? What belongs to the victim that should be present? Looking at the situations presented above, we note something peculiar: it is the victim's *physical body* that should be present. The body is used as a means, if, for example, the bodily integrity is violated (e.g. meat production, experiments, organ transplantation, bodily manipulation), if there is a sexual act with the body (e.g. rape, harassment), if the body is forced to do something (e.g. slavery), if the body is forced to be somewhere (e.g. in a cage), if the body is photographed or viewed (e.g. nude photography without consent, violations of bodily privacy) or if the body has an economic price (e.g. trafficking).

What is the body?

If the victim's body plays a central role, we have to ask the question: what is the body? One rather artificial definition of a body is: the composition of all living cells with the same DNA that are connected to each other. The artificiality makes this definition less suitable in ethics. We can also ask the question what about artificial limbs or tools that extend the body? Those extensions are strictly speaking not part of the body because the person does not have an internal representation of those extensions. Having an internal representation might be a morally relevant condition for something to belong to the body of a person. From the early stages of development, a subject creates an internal representation of his/her body: s/he learns what is part of her body and what is not.

However, there are different kinds of internal representations: I can say that this arm belongs to me, because I can autonomously move it, or because I can feel it. If I have sensations (if I can feel for example pressure, temperature and pain) in something, that thing belongs to my body. A body can be defined as those things of which someone has sensations or internal representations (think about the representations in the motor cortex and somatosensory cortex).

If internal representations – and especially sensations – are important: what about paralyzed or anesthetized limbs? As the mere means principle is related to the notion of bodily autonomy, we can say that those limbs belong to someone's body if that person still believes they belong to his/her own body. In other words, we should respect what the individual believes is part of his/her own body. This belief can either be a conscious cognitive state, or can be the internal representation itself.

But there is more. What about your gut bacteria? Or internal parasites? Suppose you are infected with a (rather harmless) parasite. I want to cut you open in order to do important experiments on the parasite. Or I want to kill you to use your gut bacteria for some important purpose. Even if strictly speaking those bacteria and parasites do not belong to your body, I still violate your basic right, because I transgress a bodily boundary. We could say that, broadly understood, everything that can only be accessed by transgressing something that has sensations, belongs to someone's body. You have sensations in your belly, so I cannot cut your belly open.

Another way to look at the problem of the use of parasites and gut bacteria is the observation that those things would not be present if the victim's body was not present. In other words, if I want to use your gut bacteria, the presence of your body (defined as those things of which you have sensations or internal representations) is required. And if you do not want to be treated (cut open) that way, the conditions of the mere means principle are met. I use your body indirectly without your permission, if I cut you open to reach your gut bacteria.

The strange thing about the mere means principle is that it points to the importance of the body, but it is not (yet) able to clarify what exactly belongs to the body. Although more has to be said on this, I will not discuss it further here. Let me conclude by mentioning that a vague boundary of what belongs to someone's body does not necessarily present a problem in the construction of a non-arbitrary consistent ethic, because the strength of the basic right (see section 6.5) might also have a gradation. We could couple this gradation with the gradation of how strongly something belongs to someone's body. If a thing definitely belongs to someone's body, the mere means principle would be strongly violated when that body part is used without consent. But if it doubtfully belongs to someone's body, the violation of the mere means principle should not count so gravely.

The reference to the body also leaves the mere means principle with another very basic question.

Why the body?

What is so important about someone's body, to give it a privileged status? From a theory of property rights, we could say that the body is the only thing that a being owns completely. The body falls under the absolute competence of a person. The deontological mere means principle, with its focus on the body, corresponds with a restricted kind of propertarian libertarianism. In the libertarian theory of property rights, agents fully own themselves and can acquire property rights in external things (Vallentyne, 2012). Propertarian libertarianism states that private property is the sole source of legitimate authority. Its non-aggression principle is restricted to violations of private property. These property rights are non-negotiable (Nozick, 1974).

The mere means principle fits in a propertarian libertarianism, where people have a full property right over their own bodies. People have a full bodily autonomy. The mere means principle does not imply that external things are owned to the same degree as bodies are owned. Owning a body is much more important than owning an external object.

In slavery and trafficking, the bodies of the victims are treated as someone else's property in the legal or economic sense. They are merchandise. This property status is not respectful, because only the victims themselves possess their own bodies. The victims do not have to be aware of this property status. For example selling babies is immoral, even when the babies do not understand the notion of private property. According to animal rights activist Gary Francione (2000), we should also abolish the property status of animals. So we should not be allowed to buy and sell animals (e.g. buying a pet from a breeder), even if those animals (like babies or mentally disabled humans) cannot be aware of their property status.

An example from economics: paying taxes versus forced labor

According to Nozick's libertarianism, raising taxes is in some way comparable to slavery: the state appropriates a part of the work and time of persons, without their consent (Nozick, 1974, ch.7). But if a libertarian restricts absolute property rights to only the body, taxation is no longer impermissible. Raising taxes is possible (even if the presence of the tax payer is necessary to raise the tax, and even if the tax payer did not give permission), because money is not completely owned by a person. Money is not a direct product of only the body of the worker.

The worker not only uses his/her own body, but s/he also uses something external. A farmer's manual labor belongs to the farmer. But the mere means principle allows for an assertion that the soil used by the farmer is not completely owned by the farmer. The soil, and everything else that is external to the bodies of persons, belongs to society. So the state can say to the farmer that if the farmer wants to use something external to his body, the state (society) has a right to interfere to the benefit of society. Hence, a part of the harvest can be given to society, in order to benefit the total good of society or the well-being of the worst-off individuals, i.e. for distributive justice. If you don't want to give away a part of your harvest, fine, but then you are not allowed to use something that is external to your body and that belongs to society.

In this sense, we can clearly distinguish taxation from slavery. Forcing someone to do labor is not allowed, because in that case the body of the slave plays a central role. It is not allowed to force someone against his will to use something external to make a product. But if persons themselves want to use something external to make products, society has a right to interfere by taxation.

If external things can be owned completely by persons, as in propertarian libertarianism, utilitarians have no grounds at all to improve well-being, and there is no

space for distributive justice. But if absolute property is restricted to the body, as in the mere means principle, utilitarians can still use taxation to improve well-being and a (Rawlsian) system of distributive justice is still possible to some degree. Hence, the only playing field for consequentialists (such as utilitarians, egalitarians, prioritarists) who want to improve or equalize well-being exists when people want to use something external to their bodies.

The deontological mere means principle takes a position between utilitarian/egalitarian/prioritarian consequentialism and propertarian libertarianism. The latter says that people can have an absolute property right over everything, the former says that there are no absolute property rights, not even over the own body. According to a consequentialist, people can not only be forced to pay taxes, but can also be forced to let their own bodies be used. We have a duty to help others in need, by paying taxes to help the poor. But the utilitarian can go further by claiming that we also have a duty and should (even without our permission) donate our blood or a kidney. The mere means deontologist can respond that such donations are morally good but not obligatory or enforceable. They are ‘supererogatory’ (good but not obligatory) because our blood and kidneys are completely owned by us as parts of our bodies so we can decide what happens to them.

6.3.2 Conclusion

The above two conditions give us a fairly precise, clear and nuanced picture when a basic right is violated. If the agent causes harm to a victim, violating his interests in a way that the victim does not want, and if the presence of this victim’s body was required in order to reach an end, then the victim is used as merely a means for someone else’s ends. We see a dual role of the presence of the bodies of the agent and the victim: the presence of the body of the agent is a necessary condition in the causation of the harm, and the presence of the body of the victim is a necessary condition in achieving the goal.

In the next section I will argue who gets the basic right. There are a lot of beings, each with different levels of complexity and interests. So giving them all an equal claim for this basic right will be difficult. There is a gradation in complexity and interests, and there is also a gradation in someone’s ends. Could it be possible to make a coherent picture by coupling those two gradations? We will see that the questions “Who gets the basic right?” and “What are the ends?” are related to each other.

6.4 Who gets the basic right?

Looking at the consequentialist QMM-principle, it was easy to see to whom the principle applied: all beings who have a well-being, i.e. all sentient beings. I also argued that the principle of tolerated choice equality naturally applies to all sentient beings, because we can only feel empathic concern for sentient beings. So we might think that the basic right principle also applies to all and only to sentient beings. But we have seen that this principle is not derived from the veil of ignorance, and it is not based on empathy.

The two conditions of the mere means principle indicate two criteria for granting someone a basic right. Each of the two criteria has a broad and a narrow formulation. The first criterion refers to the presence of the body, so the being should have a body. More narrowly formulated: the being should have an internal representation of his/her own body (it should know where its body ends and the environment begins). The second criterion refers to the interests of a being, so the being should have interests. More narrowly formulated, the being should be able to want something. In this interpretation, the being should be sentient in the sense that it has a well-being composed of positive and negative feelings related to (dis)satisfaction of preferences. Those feelings and emotions indicate what a being wants. This criterion can be narrowed further by requiring higher mental capacities for autonomy or rationality.

As the basic right is based on respect, there is a second way to solve the question who gets the basic right. We can ask who or what earns respect? My guess is that respect is connected to something complex and vulnerable. There are different complex and vulnerable things in the universe, such as living beings and sentient beings. These beings are characterized by having complex interests. Cars or stones do not have complex interests, because they don't even act to protect their interests. They can have an interest not to be broken, but that is a trivial interest. We might say that complexity in interests is related to respect. And as we have seen, respect means that we should not violate someone's basic right. Now, rights are nothing but devices to protect interests. So it is not farfetched to couple the notion of interests with the basic right. How can we do this in a natural way?

First, we observe that there is a gradation of complexity in terms of a gradation of interests (needs). Roughly speaking we have *non-living objects* with only trivial interests and low complexity. *Living beings* have complex interests (to eat, to live,...) and they have a high complexity (DNA, metabolism,...). But some living beings can perceive their environments, or respond to their environments in even more complex ways, because they have nervous systems that allow them to have inner, neural representations of their bodies and environments. Although they are unconscious (like robots), these sensorineural, perceptive or *responsive beings*, such as invertebrate

animals, have even more complex interests and they have complex reactions towards them.

But some responsive beings have more: a central nervous system that generates a perceptual consciousness. They are subjectively aware of their environments and bodies. The representations of their environments and bodies are accompanied with 'qualia' (Byrne, 2010), the subjective, private, direct, conscious experiences. Together with qualia, a sentient being has a focus or special attention towards an object (Ramachandran & Hubbard, 2001). For example: the feeling of touch in my fingertips only happens when I focus on my fingertips. Just before I paid attention to this feeling of touch, I was not aware of it. There was an unconscious neural activity (no anaesthesia), comparable to what responsive beings might experience. But only after I focused on my fingertips, it became a conscious experience or 'quale' of touch. This focus or attention is important in the conscious experience, and it might be possible to see this in the behavior of some animals, because the focus decreases the awareness of other things. For example, a cat focusing at his prey is no longer paying attention to other things. Or a fish (e.g. a trout) injected with a venom becomes preoccupied with the pain, so that it pays no heed to a threat coming towards him (EFSA, 2009). These are indicators that those animals have qualia, because they are analogous to our behavior when we have qualia and focus. Now, qualia are often neutral. I don't feel an urge to avoid touching books. The touch of a book has no influence on my will. But other qualia are affective in nature; they are evaluated as being positive or negative. For example, the feeling of a needle in my finger generates a quale that I wish to avoid. This quale of pain generates an urge in me. Those affective or evaluated qualia are the positive or negative feelings and emotions such as pain, fear, distress or joy. This is where well-being comes into play. These feelings are related to interests or needs, they are nothing else but subjective experiences of (un)satisfied interests. Fear indicates that the need for safety is not satisfied, pain indicates that the interest of bodily integrity is violated; frustration may indicate a need for freedom. Responsive beings who have evaluated (affective) qualia are called *sentient beings*. They are subjectively aware of their interests, so they not only have interests, they not only react to them in complex ways, but they can also subjectively feel them. These are the beings that have a subjective well-being, so things subjectively matter to them. They *want* things. Responsive beings with only unconscious experiences or neutral qualia, have no well-being, because the well-being is composed of evaluated qualia that are positive (joy,...) or negative (pain, frustration,...). These beings do not want anything.

Finally there are the *rational beings*. These are sentient beings with a self-consciousness and rational agency. They not only have complex interests, they not only react to them, they not only feel them, but they know and understand them. These beings have the most complex emotional lives, with a future perspective, dreams and

projects. These rational beings not only want things, they are also able to give informed consent.

The above distinction between rational, sentient, responsive and living beings can offer us some extra degrees of freedom to construct a consistent ethic that best fits our moral intuitions. We clearly have a gradation of complexity of beings. Now, looking at the definition of the basic right, it refers to the use as merely a means to someone else's ends. But the ends also have a gradation. There is a difference between luxury and vital needs. So it would be very natural to couple the gradation of complexity in interests to the gradation of the ends. Let's look at this gradation in ends in more detail, from luxury needs to survival ends.

Luxury: these are needs that have a positive contribution to someone's well-being when satisfied, but these needs are created by society. We can create circumstances where these needs no longer need to be satisfied in order to have an increase in well-being. Luxurious needs are volatile, relative and variable. Examples are fashion, social status symbols and needs created by commercial advertisements.

Basic needs: these are needs not required in order to stay healthy and alive, that have a positive contribution to someone's well-being, are stable and not determined by society. Examples are social contact, knowledge and recreation.

Vital needs: these are needs that need to be satisfied in order to stay alive and healthy, such as medicines and health care (e.g. new organs for patients with an organ failure).

Survival ends: these are vital needs that are not only important for individuals, but for biodiversity as well (e.g. survival of species). Examples are food, water, air, sexual activity (procreation) and motion. In part 3 (Chapter 10) I will discuss the predation problem, whereby we will see that there is a morally relevant distinction between survival ends and merely vital needs. Vital needs are characterized by one criterion: necessity. Survival ends, on the other hand, are characterized by three criteria: natural, normal and necessary. Natural means that the behavior is directly developed by evolution (genetic mutations and natural selection), and as biodiversity is defined by everything that directly evolved from evolution, natural behavior contributes to biodiversity. Natural plus normal means that the behavior is natural and happens a lot, and therefore contributes a lot to biodiversity. Natural plus normal plus necessary means that much biodiversity will be lost when the behavior stops. Eating food is natural, normal and necessary. Organ transplantations or medical experiments are necessary, but not natural or normal. Therefore, for a patient in the hospital, new organs or medicines can be a vital need but not a survival end. In summary, survival ends are in some sense stronger than vital needs. The difference between survival ends and vital needs is related to the moral value of biodiversity, which is threatened if survival ends are not satisfied.

The following figure represents the coupling of two gradations: complexity in interests and ends. What we see is that our approach contains the Kantian idea that rational beings are never to be used as merely a means. But we extend this basic right to other beings. Doing this makes our theory more coherent with some moral intuitions. The first intuition says that mentally disabled humans (non-rational beings) are not to be used for vital, basic and luxury needs. The second intuition is that it is self-evident to couple the basic right with the notion of interests, because rights are devices to protect interests. The third intuition is that it is self-evident to couple the complexity in interests with respect for that being, and to interpret respect in terms of the basic right not to be used as merely a means. The fourth intuition says that it is self-evident to couple gradations with each other, and the formulation of the basic right in terms of use as means for ends serves perfectly for such a coupling. This coupling immediately solves the question of who gets the basic right.

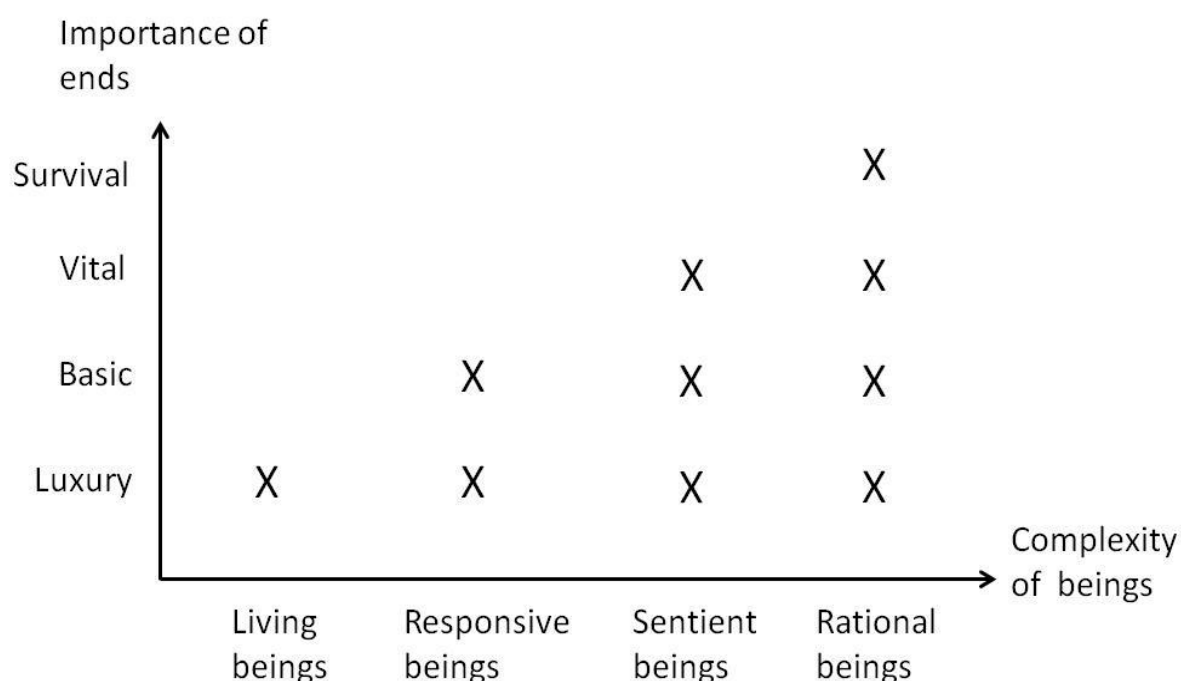


Figure 9. The coupling between ends and complexity. An X means that the being has a right not to be used as merely means for the respective ends. For example: it is not allowed to kill and use a living being for luxury needs. Rational beings are never to be used.

Looking at the above figure, we get four ethical principles.¹¹

1) All non-responsive living beings (e.g. plants and living cells) have an equal claim to the basic right not to be used as merely a means for our luxurious needs. This is a reflection of a biocentric (Taylor, 1989) or deep ecology ethics, which implies sobriety, no commercial advertisements and no status consumption.

2) All non-sentient responsive beings (e.g. invertebrates) have an equal claim to the basic right not to be used as merely a means for luxury and basic needs. We are allowed to use them for vital needs (e.g. experiments). Eating animal products (from both sentient and non-sentient animals) is not a vital need for us, because we can live healthy with a well-planned vegan diet (ADA, 2009). So eating animal products is not allowed when it is not a vital need or a survival end.

3) All non-rational sentient beings (e.g. vertebrate non-human animals and mentally disabled humans) have an equal claim to the basic right not to be used as merely a means for vital, basic and luxury needs. Experimenting on animals or using them for organ (xeno)transplantation would not be allowed. But eating animals is allowed when it is a survival end, as we will see in the section on predation in part 3. Predators (and some indigenous people) are allowed to eat meat, because they became dependent (by evolution) on other animals in order to survive. It's a survival end, because biodiversity will be lost if all predation was prohibited. Of course they are only allowed to eat animals until feasible alternatives for them are found.¹²

We have to add that sentient beings are beings who developed the capacity to feel and have not yet permanently lost this capacity. This is relevant, because we are not allowed to use sentient beings when they are asleep or temporarily unconscious. Even when they can temporarily not feel anything, it is not respectful to use them as merely a means.

4) All rational beings (mentally healthy human adults and children) have an equal claim to the basic right never to be used as merely a means, for no ends at all. Eating rational beings is never allowed, not even for survival ends. We should protect rational beings from predators if we can.

Of course violations of the basic right are allowed when the QMM-principle is very strongly violated. The basic right is not absolute, because our moral intuition says that it would be inefficient to let thousands of people die simply because we don't want to

¹¹ It is actually a continuum of principles, because not only are there four types of beings, but a spectrum of beings with gradually increasing complexity. There is also a spectrum of ends, because there are no sharp boundaries between e.g. luxury and basic needs. We can couple these two spectra.

¹² In some situations, killing and eating mentally disabled humans might equally be permissible (otherwise it would be speciesist). But the tolerated partiality principle says that it is equally permissible to prefer eating non-human animals instead of those disabled humans.

violate the basic right of one individual. We have a small but non-zero need for efficiency, just like we had a small but non-zero need for efficiency in the context of the QMM-principle. This brings to the next section.

6.5 How strong is the basic right?

If the mere means principle is an absolute principle, it would correspond with a basic right of infinite strength: a constraint that can never be passed. But a lot of people have the intuition that the mere means principle should not be absolute. Looking at the bridge trolley dilemma, we can say that the basic right is stronger than at least five times the right to live. But in the end, a lot of people have a non-zero need for efficiency in well-being: the death of millions of people might surpass the basic right of one individual, because the loss of utility (well-being) becomes too big. This means that the moral force (the strength) of the basic right is lower than the moral force of a huge amount of well-being.

As in principlism (Beauchamp & Childress, 2001), some intuitive balancing between the mere means principle and the consequentialist principle (QMM-prioritarianism) is required. It is a balancing of 'moral forces', comparable to the forces in physics: gravity is much weaker than electromagnetism, so gravity can surpass electromagnetism only when gravitational masses are very big and electric charges are relatively small.

Well-being can sometimes surpass the basic right, or in other words: the ends can sometimes overrule and justify the means. Hence, a first advantage of a finite strength (a non-absolute principle where the basic right has a finite weight) is that our intuitive need for efficiency can still be met. But the strength of the mere means principle can also depend on some other variables: how much harm is caused to the victim? How strongly does the victim want to avoid the treatment? How much of the treatment is disliked by the victim? How much consent does the victim give to the treatment? How strongly does something belong to someone's body? And what are the mental capacities for autonomy of the individual? These variables should be included in the intuitive balancing.

As a consequence, a second advantage of a basic right with a finite strength is that it allows a coupling between different gradations: the strength of the basic right can be correlated with how much harm is caused, how much the victim does not want the treatment, how strong something belongs to the victim's body and how directly the body is used.

Finally, a third advantage of a finite strength is that it allows for a consequentialism in deontological rights: a world where the basic right of one individual is violated is

better than a world where the basic rights of two individuals are violated in a similar way. If the strength of the basic right would be infinite (absolute), we do not have this property (as two times infinity equals infinity).

So we have three advantages of a non-absolute mere means principle: compatibility with a need for efficiency, coupling with gradations and consequentialism of basic rights violations. One disadvantage is that a non-absolute principle requires an intuitive balancing. This intuitive balancing should be done by all moral agents in a democratic way¹³ (see section 4.6.11): we should take a democratic average of the moral force of the basic right. This can also be done mathematically, see the next intermezzo: the welfare function can include R-parameters that measure the violations of the basic right. The average is democratic in the sense of being unweighted (taking an unweighted average of the R-parameters in the welfare function): all preferences (intuitive balancing of the moral forces) of all moral agents count equally.

Just as physics is not inconsistent when electromagnetism counteracts gravity, so is ethics not necessarily inconsistent when the mere means principle counteracts the QMM-principle. Inconsistency might occur when the balancing between these two principles is arbitrarily applied in different situations. As if gravity arbitrarily gains strength even when masses remain equal. The strength of forces in physics should conform to universal laws.

An example of an inconsistent balancing of moral forces occurs in situations of discrimination, where the moral force of the basic right of one individual is estimated to be stronger than the basic right of another individual who should have an equal moral status. As we will see in the next chapter, speciesism is a kind of discrimination that is a moral illusion. Such moral illusions can create biases in the balancing of moral forces by moral agents. Moral agents who are vulnerable to moral illusions might have inconsistent estimates of the strength of the basic right of different individuals. This is important in e.g. discussions on the use of sentient beings for medical experiments: animal researchers, as moral agents, might have an illusory bias towards using non-human animals, as if the basic right of those animals is weak compared to well-being (of humans). However, they would have very different estimates for the basic right of some mentally disabled humans: those humans have a much stronger basic right although

¹³ Note that experiments demonstrated that a moral agent can have differing intuitive estimates of the strength of the basic right, depending on some irrelevant circumstances and cognitive biases: the influence of induced feelings of disgust and humor (Greene, 2008), the framing of the description of a situation (Petrinovich & O'Neill, 1996; Sinnott-Armstrong, 2008; Lanteri et al., 2008; Ray & Holyoak, 2010) or the order in which dilemmas are presented (Liao et al. 2011; Schwitzgebel & Cushman, 2012; Di Nucci, 2012). Therefore, a moral agent would have to agree that his/her intuitive estimate lies in a certain range, so s/he should be very flexible and tolerant towards changes of the strength of the basic right within this range.

there is no consistent justification for that difference in estimated strength between those humans and animals. Hence, the democratic averaging of the moral force of the basic right is only valid if the moral agents do not have moral illusions such as speciesism.

The basic right principle not only overrules the QMM-prioritarian principle in a lot of situations, but also overrules the tolerated partiality principle (tolerated choice equality) in all situations. The tolerated partiality principle is too weak and can never trump the basic right. To see this, take another look at the burning house dilemma: your child or an unknown child? If you save your child, you are not using the other child as merely a means. Now let's look to a dilemma that is structurally very similar to the burning house dilemma. Suppose you are a surgeon and in the hospital is your child and an unknown child. Your child needs a spleen, the other a liver in order to survive. For the moment, you can keep them both alive for some days by administering a drug. If you do nothing, both will die, as in the burning house. However, you could stop giving the drug to the other child, so that child dies. Then you could use his spleen to save your child. In this situation, people are very reluctant to say that the surgeon is allowed to let the other child die in order to use his organs. So, the ethical principle that you may prefer to save your child from the flames in the burning house does not imply that you are also allowed to save your child in the hospital, by killing (or letting die) another child in order to use his organs for a transplantation.

6.5.1 Intermezzo: combining the basic right with the prioritarian theory

This intermezzo is meant for the people who have read the mathematical section in the chapter on the QMM-theory. The question is: Can we incorporate this new basic right principle in our mathematical formulation? We suggest the following possibility: apart from their values of life, all individuals have a 'basic right parameter' r_i^X , which is zero if the basic right of individual i in situation (world history) X is not violated, and very large if her basic right is violated. The quantity r_i^X can also take different values, depending on what the 'ends' are and how seriously someone's will is violated. With these basic right parameters, we add a new term to the welfare function: the basic right function

$$R(X) = - \sum_i^N r_i^X.$$

When the basic right of person i is violated, the value r_i^X should be very large, but not infinite: there might be a threshold value, above which one prefers the consequentialist outcome.¹⁴ For example: when there are, say, a billion people on the main track of the trolley dilemma (all threatened), one might be tempted to push the heavy man from the bridge in order to save those billion people. Another reason why r_i^X should not be infinite, is because it is difficult to count with infinities, as infinity plus infinity equals infinity. So when we have to choose between situation X where one basic right is violated and situation Y where the basic rights of two persons are violated, we should be able to conclude that the situation X is preferred over Y .

There is a basic right equality if $r_i^X = r_j^Y$ when the use as a means of subject i in situation X is similar to the use of subject j in situation Y (i.e. for the same kind of ends).

Having said this, we can now write the new mathematical formulation of the QMM-principle extended with the deontological rule of the basic right. In the intermezzo in section 4.6.8, I presented a welfare function that depends on the momentaneous well-being and a connectivity function. Similarly, we should write the basic right term as $r_{\pi(t)}$, which represents the level of basic rights violations of the momentaneous mind π at time t . I will also include the ex-inter approach to lotteries, i.e. using expectation values, written by the brackets $\langle \rangle$, that calculate the probability weighted average. Now we can write a moral weight, which is composed of two terms: the welfare function of consequentialist ethics and a basic right function of deontological ethics. This moral weight function reads:

$$\begin{aligned} M(\mu, c, r) &= W_{QMM}(\mu, c) + R(r) \\ &= \frac{v_+}{v_+ + v_R} \left(\left\langle \frac{1}{v_F \Delta t} \int_0^{\Delta t} \sum_{\pi(t)=1}^{N_F(t)} \left(\left(\int_0^{\Delta t} \sum_{\pi'(t')=1}^{N_F(t')} c_{\pi(t), \pi'(t')} \mu_{\pi'(t')} dt' \right)^+ \right)^p dt \right\rangle \right)^{\frac{1}{p}} \\ &\quad + \frac{v_-}{v_R} \left\langle \frac{1}{v_F \Delta t} \int_0^{\Delta t} \sum_{\pi(t)=1}^{N_F(t)} \left(\int_0^{\Delta t} \sum_{\pi'(t')=1}^{N_F(t')} c_{\pi(t), \pi'(t')} \mu_{\pi'(t')} dt' \right)^- dt \right\rangle \\ &\quad - \left\langle \int_0^{\Delta t} \sum_{\pi(t)=1}^{N_F(t)} r_{\pi(t)} dt \right\rangle \end{aligned}$$

This moral weight function can be considered as the standard model of ethics, just as a Lagrangian represents the fundamental quantity in the standard model of physics

¹⁴ The strength of someone's basic right (the value r) can depend on both the number of lives saved and the number of lives at risk. If for example a trolley is about to kill N number of people (N lives at risk) and pushing a heavy man in front of the trolley saves M number of people, the permissibility of pushing the man can depend on M as well as on the ratio M/N . An experimental study found that the higher N , the higher M needs to be to make the act permissible (Rai & Holyoak, 2010). This dependency on the ratio M/N passed a coherence test.

(Weinberg, 1996). The moral weight combines the consequentialist welfare ethic with the deontological mere means principle. We should not try to maximize the moral weight directly. Instead, we should use the method of rule universalism: derive those rules that, under universal compliance, would maximize the above moral weight. Those rules should be followed by all moral agents who are capable of following them.

In the next section, we will see an extended mere means principle, which generates the tolerated partiality principle. Hence, the basic right function $R(r)$ in the above expression can refer to the extended mere means principle, which says that we should not *use* nor *consider* someone as merely a means. As we will see in the next section, not considering someone as merely a means implies that we should allow for some level of partiality.¹⁵ If we consider a person as merely a means, i.e. if we do not allow that person to be partial whereas we should allow such degree of partiality according to the extended mere means principle, then a negative r -term is added to the moral weight. Or in other words: when you want to be partial to a degree that should be universally permissible, and when I prohibit you to be partial in that way, then a negative term is added to the equation.¹⁶

In this way, all three principles are included in the above mathematical expression: QMM-prioritarianism, tolerated partiality and the mere means principle. With this expression of the moral weight, including the tolerated partiality principle in the basic right function, we can get a very rich ethical system that is compatible with a lot of moral intuitions that a lot of people have.

6.6 The extended mere means principle

This section discusses an interesting relation between the basic right (mere means) principle and the principle of tolerated partiality. The mere means principle can be extended, from using someone as merely a means to considering someone as such. This extension can help us to explain the differences between doing and allowing as well as positive and negative duties. I will demonstrate that making these differences, using the

¹⁵ As we will see, the extended mere means principle does not allow for levels of partiality where someone's basic right is violated. Saving your child by killing another child and using its organs for transplantation, is not allowed. Such behavior would be too partial.

¹⁶ If you do not want to be impartial, I cannot prohibit you to be partial. The only thing I can do is try to convince you to behave more impartially. Suppose I convinced you such that you want to be impartial. Then I do not consider you as merely a means when I want you to be impartial. In that case, no basic right is violated.

extended mere means principle also generates an argument to justify the tolerated partiality principle.

6.6.1 Doing versus allowing

Many people have the intuitive moral judgment that doing harm is worse than allowing a similar level of harm (Kagan, 1989, p. 94). Pushing a child in the water to kill him is worse than not saving a drowning child. However, this often heard drowning child example is not a real moral dilemma: it is not a choice between pushing versus not saving. So this example is misleading. A better example would be the following, ‘switch trolley dilemma’. A runaway trolley is about to kill one person on the main track. You can turn a switch and send the trolley to a side track, where it will kill another person. It is a dilemma, because you now face a choice between actively turning the switch versus doing nothing. In this dilemma it becomes less obvious that turning the switch and killing the person on the side track is worse than allowing the person on the main track to die. A lot of people say that it is permissible to turn the switch (especially if the person on the main track is your child) (Hauser et al. 2008).

Now imagine there were three people on the main track, and the person on the side track is someone I know. I do not want to kill this person on the side track, so I let the three people on the main track die. You could say that I had a duty to turn the switch, because one dead person is better than three dead people. But if you would say that to me, you would *consider* me as merely a means to the ends of the three people. You would not literally use me as merely a means, but according to your judgment, my presence was required to save the three people, and I would have to do something (turning the switch) I do not want.

If you are not allowed to judge me for not turning the switch, it *appears as if* allowing the three people to die is not worse than killing one person. Hence, if we extend the mere means principle, from using someone as merely a means to considering someone as such, we have coherence with a deontological rule of doing versus allowing. The extended mere means principle generates an *apparent*¹⁷ difference between doing and allowing. This difference corresponds with a counterfactual account of the doctrine of

¹⁷ Note that the extended mere means principle merely says how I am not allowed to judge or consider you. This does not imply that you do not have certain duties. You still might have a duty to turn the switch to save the three people. That duty is compatible with my duty not to judge you if you do not turn the switch. Even if I am not allowed to consider you in a certain way, not much follows from this how you are allowed to act. We have to distinguish primary duties (how to act) from secondary duties (how to judge actions). Nevertheless, we could say that my secondary duty (not to judge you) counts as a justification for your (lack of) primary duty.

doing versus allowing (see Howard-Snyder, 2011), which says that the presence or absence of the agent is morally relevant: a 'doing' requires the presence of the agent. If an upshot would not have occurred if the agent had been absent from the scene, the agent was not positively relevant to the upshot (see Kagan, 1989, p. 94).

The difference between doing and allowing also corresponds with a difference between positive and negative duties. A positive duty is a duty of beneficence, where the presence of the agent (the helper) is required in order to benefit someone. A burning house dilemma exhibits positive duties: the helper can only save someone and cannot cause harm to someone (when the helper does nothing, s/he allows harm). A negative duty of non-maleficence (the no-harm principle) does not require the presence of the agent: if the agent is not present, the no-harm principle is trivially satisfied because the agent cannot cause harm when s/he is absent.

In situations of negative duty, we can judge someone for violating his duty of non-maleficence, without considering him/her as merely a means. However, if you do not want to help someone, and if I claim that you violate the duty of beneficence, I would consider you as merely a means. Therefore, violations of positive duties are considered less bad (more tolerable) than violations of negative duties. In the next section, I explain why partiality not only is, but actually *should* be more tolerated in positive duties than in negative duties.

6.6.2 Tolerated partiality and imperfect duties

In the previous Chapter 5, I discussed the principle of tolerated partiality, which violates the consequentialist prioritarian principle. But also the mere means principle (the basic right) violates the consequentialist principle. There is a subtle connection between the partiality principle and the extended mere means principle.

Positive duties are imperfect duties, in the sense that, while we are not required to live up to them at all times, these duties are deserving of admiration. Helping others is an imperfect duty, because there is a whole range of possible levels of assistance that one could give. Perfect duties on the other hand can and should be respected at all times (for example the duty not to use someone as merely a means).

Looking at the intuitions of a lot of people, we appear to tolerate partiality in positive, imperfect duties, but we are not so tolerant towards partiality in negative, perfect duties. The reason why partiality is, and in fact should be, tolerated in imperfect duties has to do again with the mere means principle in its extended version: do not consider someone as merely a means.

Consider a double trolley dilemma, which is similar to the burning house dilemma (choosing between one person you hold dear versus three unknown persons), but has a better structure to discuss the issue of partiality in imperfect duties.

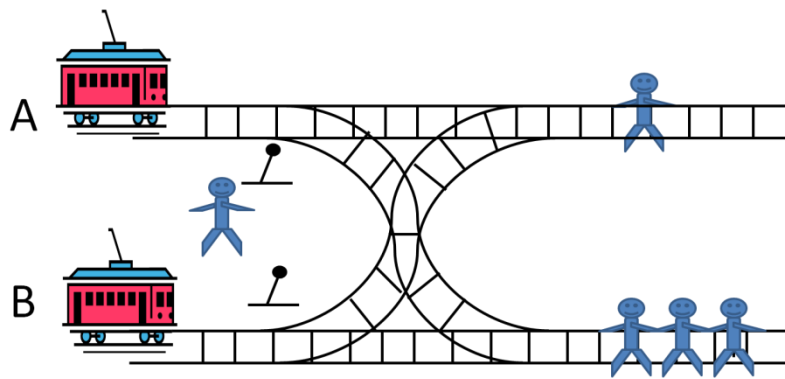


Figure 10: the double trolley dilemma

There are two tracks, two trolleys and two switches. Trolley A will kill your child on track A, trolley B will kill three unknown children on track B. You can only run to one switch. If you run to switch A, trolley A will move to track B, and your child will be saved. If I say that you must run to switch B in order to save the three children, I would consider you as merely a means: your presence is required to save the children (the ‘means’ part), and you would have to do something you do not want (the ‘mere’ part). I am not allowed to consider you as merely a means, I am not allowed to judge you, and therefore I should tolerate your partiality towards your own child. Therefore, this partiality is tolerated. The same goes for the situation when trolley A is not present: you do not have a duty to kill your own child by turning switch B in order to save the three children.

The question remains whether partiality should be tolerated when trolley B is not present. When you run to switch A in order to save your child, can I blame you? In this case, three children are harmed by your action (you cause harm; it is not an allowing to die). If I would say that you should not have turned switch A, I am not considering you as merely a means, because your presence is not required in order to do what I want (saving the three children). According to my intuition, I would not easily tolerate a partiality where you do more harm, but I do not know what the intuition of most people is in this situation.

In a previous Chapter 5 we encountered a slightly related trolley problem involving three tracks. Imagine a trolley moving towards five people on the main track. They will all die, unless you send the trolley to a side track, where it will kill someone you hold dear. But this time you also have a third option, sending the trolley to a third track, killing two people. You are allowed to save the five people on the main track. But are you allowed to send the trolley to the third track, killing two unknown people instead of the one beloved person on the second track? If you send the trolley to the third track, I am not allowed to judge you by saying that you should have send the trolley to the second track. That is because such a judgment would still imply me considering you as merely a means (because your presence was necessary to turn the trolley away from the five people on the main track, and killing your beloved one is not what you like).

Therefore, the extended mere means principle implies that you are allowed to cause more harm than you could have avoided. You are allowed to kill two unknown people instead of one beloved person, violating the consequentialist prioritarian principle, only if your action is the result of saving even more people (i.e. saving the five people on the main track).

To conclude: the extended mere means principle can explain what kinds of partiality we *should* tolerate. Not tolerating some partiality would imply *considering* someone as merely a means, which is immoral.

6.6.3 The asymmetry of procreational duties

The ‘asymmetry of procreational duties’ (Narveson, 1967; Mulgan, 2006; McMahan, 2009) says that we do not have an obligation to give birth to happy children (out of the interests of those children), but we do have an obligation *not* to give birth to children when we know that the lives of those children will be miserable. Think about the problem of parents knowing they have a genetic defect which means that their potential child will be seriously handicapped. Or think about animals raised in the livestock industry. Those animals are bred for their high productivity, which often means that they suffer from serious physical problems (e.g. big udders, lower immunity, deformations).

As we have seen in a previous chapter on population ethics, the welfare function derived behind the veil of ignorance shows a threshold of well-being for high population sizes: if a newborn sentient being gets a lifetime well-being below some power-averaged value, it lowers the welfare function. I argued that a deontic permission still allows for the procreation of such beings. But what if a newborn sentient being will get a lifetime well-being above the threshold, such that the welfare function increases? Do the parents have a duty to procreate in this case? The extended mere means principle says that we cannot judge parents who do not want to procreate, even though their future children would increase the welfare function. That is because we cannot consider those parents as merely a means (as breeding machines), doing something that they do not want. In other words: the asymmetry of procreational duties is coherent with our deontic extended mere means principle. Other possible solutions to explain the asymmetry (as in e.g. McMahan, 2009) are not needed.

In summary, there are two relevant levels of lifetime well-being and two different principles that make procreation permissible (i.e. neither a duty nor a prohibition).

- 1) If a potential being would have a lifetime well-being above some positive threshold such that the welfare function would increase if the potential being gets born, we do not have a duty to procreate, because we cannot consider a woman as merely a means to increase the welfare function. The mere means principle implies that the woman is

allowed to be partial towards her own preferences, because the woman can decide what happens to her body.

2) If a potential being would have a lifetime well-being below that positive threshold but still above zero, i.e. if the potential being would still have a life worth living, then we definitely do not have a duty to procreate (because that would lower the welfare function), but we have a deontic permission to procreate. We may procreate and give birth to that potential being if we want to. The biodiversity principle (the 3-N-principle, to be discussed below in section 10.4) implies that procreation is still allowed.

3) If a potential being would have a lifetime well-being below zero, i.e. a life not worth living, we have a duty not to procreate.¹⁸

6.7 Application: the least harm principle and vegetarianism

The basic right principle has a lot of implications. It might solve an important objection against vegetarianism. Davis (2003) argued that an omnivore diet, killing and eating big grazing animals (e.g. cows) would cause less harm to sentient beings compared to a vegan diet. A vegan needs a crop field, so it might be possible to count how many animals (e.g. mice) die by accident using this crop field. Suppose that five mice are accidentally killed in the harvesting process to produce the same amount of nutrients as the omnivore's diet where only one cow is used. Then the omnivore causes less harm than the vegan.

Whether a vegan diet causes more suffering or death is a scientific question that is strongly debated due to lack of good data. At least Matheny (2003) and Lamey (2007) criticized the argument of Davis. But here we can somehow avoid this issue, by introducing the basic right. The omnivore uses the cow as merely means, so the basic right of the cow is violated. On the other hand the field mice are accidentally killed, so they are not used as merely a means. If the basic right trumps the right to live, a vegan diet remains more ethical.

¹⁸ Although I am not so sure about this third rule. It might be the case that a lot of animal species (especially the species with a reproductive strategy of so called *r*-selection) give birth to short miserable lives that are not worth living: a majority of those animals starve or are preyed upon and die shortly after they come into existence (see e.g. Horta, 2010c). If all those species are not allowed to procreate, a lot of biodiversity will get lost. I suggest we take much more effort and do scientific research to increase the lifetime well-being of those animals by e.g. redesigning nature.

We can compare this with driving a car. It is true that car traffic sometimes accidentally kills children. Now imagine (hypothetically) that we invent a new form of transportation, some kind of teleportation device. However, this device can only work if you kill a person and use his body to drive the teleportation device. Are we allowed to kill and use an innocent person as merely a means, in order to save more children from car traffic? I expect that most people have the intuition that using a person for the teleportation is not allowed.

Chapter 7 Summary: principles of equality and further refinements

In the previous three chapters, we encountered three ethical principles of normative ethics. Each principle corresponds with a normative ethical theory: an ethic of welfare (consequentialist ethic), an ethic of care (relational ethic) and an ethic of rights (deontological ethic). The quasi-maximin theory is the consequentialist principle of prioritarian justice. It is based on the fundamental ethical notions of impartiality (justice) and well-being. This serves as an underlying structure or backbone for the other two normative ethics. Due to some moral intuitions that violated this QMM-principle, we introduced two other universalized ethical principles that overrule the QMM-principle. The first is a weak overruling that uses elements of an ethic of care (empathy in personal relationships). It says that we are allowed to be partial to some degree, as long as we respect similar levels of partiality of others (it generates a tolerated choice partiality). The second is a strong overruling, based on a deontological notion of a basic right. This basic right not to be used as merely a means strongly trumps both the weak overruling principle of tolerated choice, and the consequentialist theory of QMM. The following scheme gives an overview of the relation of the three principles of normative ethics.

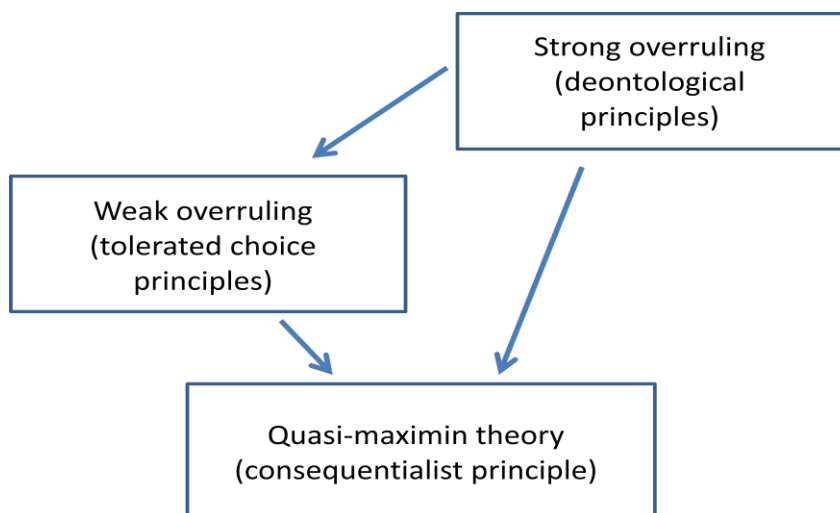


Figure 11: weak and strong overruling of the QMM-theory

All three principles contain a notion of equality. Also the universalist principle discussed in the first part of this dissertation contains a notion of equality. This leads us to four principles of equality, one formal and three material principles.

1) Formal equality of universalist impartiality: treat all equals in all equal situations equally. This is a formal principle because it does not say how we should treat everyone. This formal principle is applied to the other three, material principles of normative ethics. For example in the QMM-theory, this formal equality results in an important symmetry property of the mathematical formulation: the lifetime well-being of individuals is interchangeable.

2) Prioritarian welfare equality. This equality means that governments, professional health care and economic and legislative structures should be impartial and should strive towards a just distribution of well-being, according to the QMM-principle. It is a material equality principle, because it gives content to how well-being should be distributed. Formal equality of impartiality says that a unit of well-being counts equally for all sentient beings in all similar positions. I.e. the identity of individuals is not important when it comes to distributing well-being. But the material equality adds substance, by claiming that priority should be given to the individuals in the worst-off positions. As a result of this priority, if total lifetime well-being is constant between situations, the situation which has the most equal distribution of lifetime well-being is the best.¹ Hence, prioritarianism lies in between sum-utilitarianism (maximizing well-being) and egalitarianism (equalizing well-being).

3) Tolerated choice equality: even though we would save our own child in the burning house dilemma, we would tolerate the choice of someone else who has saved the other child. In general we should tolerate small levels of partiality, especially when personal relationships are involved. This partiality in personal relationships is central in an ethic of care.

4) Basic right equality: all beings with a same level of complexity in interests, should have an equal claim to a basic right not to be treated as merely a means to someone else's end.

With these four principles² we can get a fairly nuanced picture of animal equality, as we will see in the next part of this dissertation. The reason why we get a nuanced

¹ This was seen on the expression of the welfare function: $W = P \cdot A \cdot (1 - I)$, with P the population factor, A the average lifetime well-being and I an inequality metric.

² In the next chapter I will discuss a fifth equality principle to solve the predation problem. This fifth principle is a principle of behavioral fairness: if a zebra is allowed to eat for survival, then so is a lion. More generally: everyone has an equal right to a behavior that is both natural, normal and necessary (a behavior that contributes to biodiversity). So we end up with five principles of equality.

picture is that the equality principles correspond with our moral intuitions, and are not in contradiction with at least two other kinds of inequality.

1) Inequality of outcomes. This is strict egalitarianism, which strives for complete equality in well-being. This is not undesirable. It violates a strong moral intuition that says that inequality in well-being is permissible if it is at the advantage of the worst-off individuals. It is better to have two persons with levels of well-being 10 and 20, than levels of well-being both equal to 1.

2) Emotional inequality. We are allowed to give some preference to those individuals we hold dear. We do not have a duty to be impartial in our personal relationships. We do not have to toss a coin in a burning house dilemma, if we have to choose between saving our own child and saving another child.

Emotional inequality is not in contradiction with e.g. the universal declaration of human rights, which says that all human beings are born free and equal in dignity and rights. But in the third part of this dissertation, we will argue that it is a kind of discrimination to limit this equality principle to only humans. We will see that it is possible to extend the moral community (the 'circle of equals') to all sentient beings, claiming that all sentient beings are equal in the above four senses. This extension is necessary if we want to stick close to our strongest moral intuitions in a consistent way. Such an extension would make our theory more compatible with our strongest intuitions, compared with the current inconsistent ethics of our speciesist society.

7.1 Equality and veganism

If we extend the material principles of equality to animals, then we see that animals in the current livestock and fishery industries are maltreated in three ways.

First, the consumption of animal products is likely a violation of the QMM-principle: it is impossible to imagine that humans, if they were not allowed to eat animal products, would be worse-off than animals bred in factory farms and slaughtered in slaughterhouses. The pleasure of the taste of animal products (meat, milk, fish,...) does not outweigh the suffering of those animals. From behind the veil of ignorance, you cannot prefer a world where eating cheese is allowed. In such a world, you have a non-zero probability to end up being a dairy cow with a low value of life equal to say 3 (because you suffer in the livestock industry, and you have an early death). You also might end up being a human who is able to enjoy the taste of cheese (he has a value of life equal to say 11). On the other hand, in a vegan agriculture, this person can no longer enjoy eating cheese, so his value of life decreases a tiny bit, to say 10. But there will be no dairy cows in lower positions, so you would not have a probability to end up worse.

Being risk averse, you would prefer the vegan world, because then you do not have a probability to get a value of life equal to 3. If a cow is born, the impartial observer behind the veil prefers the cow to have a well-being of e.g. 5 instead of a lower well-being of 3 in the livestock industry. Quasi-maximin prioritarianism would therefore imply veganism.

Also the principle of tolerated partiality is violated in the livestock industry. If we tolerate the choice of a dairy farmer to use the milk of a cow in order to increase the well-being of a human who loves cheese, then we should also tolerate someone who makes the opposite choice, such as breeding women and using their breast milk to make cheese to give to animals who like breast milk cheese. This, however, we would never tolerate.

Third, the use of animals and animal products is a violation of the basic rights of animals, because these animals are used as merely a means. The bodily integrity of dairy cows is violated (by artificial insemination, forced milk production and early death) and they are treated as property.

These three different criticisms of the livestock and fishery industries should not be confused with each other. Speciesism causes serious violations of three ethical principles of equality, based on justice, empathy and respect. When applied to animal ethics, the equality principles give a complete picture of equality that extends approaches in the literature. For example, Francione (2000) only focused at the basic right (the property status of animals). This is a 'negative' approach, in the sense that it only says what we are not allowed to do (related to negative rights of not being treated in some ways). The prioritarian theory of justice also offers a 'positive' ethics, because it says something about our duty to help others (and a corresponding positive right to be helped).

The three material principles of equality do not stand on their own. They have to be combined with a universalist imperative. As we have seen, this principle of universalism implies four universalizations, two of them are particularly important: universalization with respect to the moral agents and with respect to the moral patients. The first universalization with respect to moral agents will be further discussed in the next section, where it is related to non-ideal theory (i.e. situations without universal compliance). After that, the second universalization with respect to moral patients will be discussed. As we will see, this second universalization is related to antidiscrimination and the absence of certain hierarchic dualisms.

7.2 Ideal and non-ideal theory: applying the universalist imperative

The universalist imperative says that we should not directly follow e.g. the prioritarian quasi-maximin principle on our own. This universalist imperative is a bit related to the golden rule. We can state it in different ways. For example according to the Kantian categorical imperative: Act only according to that maxim (moral rule or guiding principle) whereby you can, at the same time, will that it should become a universal law. Or: abide by those principles which we would like everyone to abide by. Or: give the good example and follow the rule that every moral being (everyone who is capable, rational and informed) should have to follow, even if no-one else does so.³

This universalist imperative reflects an unconditional commitment and we should, if need be, swim up against the stream. We should abide by those rules which are universalizable, which means that if every moral agent (who is capable and informed) should follow those rules and consequently apply them, there will be no undesirable consequences that violate one of the above principles of equality.

When choosing a rule-based action (an action based on a maxim or a guiding principle), we should ask ourselves: what are the consequences if everyone (who is well informed and able to do that action) does a similar action or follows that rule? In other words: what are the consequences in an ideal utopian world with universal compliance to the rule? If the consequences are good (if they satisfy the three material principles of equality), then we should do that action or follow that rule, even if others don't.

If we want to do an action, but we cannot find an underlying ethical guiding principle or rule for that action that can be universalized to all similar situations, we should not do that action. Actions are only permissible if you can find a justifying universalized rule that is consistent with the ethical system (i.e. a rule that does not violate a principle of the system). For example: if I want to take the train this morning, the rule "Everyone has the right to take this train this morning" cannot be applied to all persons, because that would result in an overcrowded train. But I can find another guiding principle that guides my action to take the train and that can be generalized: "Everyone has the right to take the train when there is still some place available on the train, when a fair distribution of train rides is possible and when there is no-one left who wants to occupy the free place and has an equally strong or stronger right or reason to take that train". So I can justify my use of the train by referring to this second principle.

³ This reference to rules turns the theory in a "rule consequentialism" instead of an "act consequentialism" (Hooker, 2011).

In other words: I am allowed to do an action (or inaction) only if I can find a rule or guiding principle that can be used to guide the action, given that 1) it is okay for me if this rule gets universalized (the non-arbitrariness condition) and 2) the rule is compatible with all other principles of the ethical system (the consistency condition). When I can't find a consistent, universalized guiding principle that justifies the action, the action (or inaction⁴) is not allowed.

It might be possible that the universalist imperative does not give an exclusive answer to the question what guiding principles we should act upon. If we see that two different kinds of actions or rules are compatible with the universalist imperative, i.e. if universal compliance to guiding principles A and B give the same good consequences, then we break the tie by a reality check. In reality, i.e. in a less ideal world without universal compliance, not everyone will follow that specific action or guiding principle. So we should look at the consequences of our rules and actions in the current, real non-ideal world (without universal compliance). If guiding principle A would have preferable outcomes in the current non-ideal world than principle B, we should act according to guiding principle A.

So we start with an ideal theory: deriving those guiding rules that, with universal compliance (amongst all people who are able to follow the rules), generate the best results according to our principles of equality. If there is a tie between those derived rules, we can select the best of those rules by looking at non-ideal theory, i.e. by looking at the consequences if there is no universal compliance (and in particular: if there is as much compliance as in the current real world⁵). Non-ideal situations will serve as tie-breakers.⁶

⁴ The inaction refers to e.g. not helping someone in need. When I don't help someone, I should come up with a rule that explains my not helping. If I can't find such a rule that I am willing to see universalized, I have the duty to help. If my guiding rule is simply "I never help" or something like "I never help when I don't feel like helping", I will not be willing to do the universalization, because that would mean no-one might help me when I am in need. If my rule is "I don't help at moments when I recently already helped a lot of others", I am willing to universalize this rule, so then it is okay not to help at that moment.

⁵ There are many degrees of non-compliance, so there will be many different non-ideal theories. The preferred non-ideal theory is the one that is applicable to the current real world, i.e. the one derived from the current level of non-compliance.

⁶ Parfit (2011) discussed a similar solution to the "ideal world objections" where a universalized rule might have bad consequences in non-ideal situations lacking universal compliance. According to Parfit, we should adopt new, conditional rules. For example: "Follow the rules whose being followed by everyone would make things go best, unless some other people have not followed these rules, in which case do whatever, given the acts of others, would make things go best." (p.262). This principle needs further refinements that I will not discuss here.

The prisoner's dilemma in game theory can illustrate non-ideal theory. The dilemma faces a choice between cooperation and defection. If both players in the prisoner's dilemma game cooperate, they generate the best outcomes (according to the prioritarian QMM-theory). However, if one of the two players defects, the cooperator loses and the gains for the defector increase. If both players defect, they generate a suboptimal outcome. The following table presents the possible outcomes (levels of well-being) of a prisoner's dilemma for the two players (bold type values for one player, italic values for the other).

		<i>Player 2</i>	
		<i>Cooperate</i>	<i>Defect</i>
Player 1	Cooperate	3,3	5,0
	Defect	0,5	1,1

The iterated prisoner's dilemma allows for multiple, successive rounds. Ideally, the best outcome for both players in an iterated game is mutual cooperation. But if a player who always cooperates encounters a defector, the cooperator loses. It is shown that in a non-ideal world, with defectors, often the best strategy in an iterated prisoner's dilemma is 'tit-for-tat' or 'equivalent retaliation' (Hargreaves-Heap & Varoufakis, 2004, p.191). According to this strategy, the first move (in the first round) is always cooperation. The second move (in the second round) is the same as the other player's move in the previous round. If the other player defects, you retaliate in the next round by defecting. If the other player cooperates again, you are forgiving and cooperate in the next round.

So a player has at least two good rules in the ideal world: 'always cooperate' and 'tit-for-tat'. If everyone follows tit-for-tat, the result will be continuous cooperation. To break the tie between these two strategies, note that tit-for-tat is better than 'always cooperate' in a non-ideal world. So the player can follow tit-for-tat⁷ (preferably with some level of forgiveness⁸).

⁷ Adding an exception, a rule like "always cooperate" is specified into a rule like "always cooperate, unless others don't (then follow tit-for-tat)". A criticism of rule consequentialism is that you can always further specify a rule, such that in the end you end up with an infinitely specified rule, which is equivalent to act consequentialism (Smart & Williams, 1973). Rule consequentialism collapses into act consequentialism as long as there is room for adding exceptions to the rule ("Do X unless Y"). I think that the approach of specifying rules (using exceptions) is valid and permissible. It is allowed to move closer to act utilitarianism, on the important (non-trivial) condition that one does so on a path of universalized rules, i.e. that one always refers to (specified) rules that everyone should follow. The more specified, the more complicated a rule becomes, and complicated rules have disadvantages. Where you stop along this path is up to you. As long as you can find a (specified) rule that permits your action after universalization, your action is allowed.

⁸ It might happen that the other player defects by mistake, ending up in a vicious circle or an unending "death spiral" of mutual defections. To avoid this, a good player should sometimes be a bit more forgiving, by

In real life situations, there are often more than two players who can cooperate or defect. As the universalist imperative (do what everyone should do) is related to the important idea of giving the good example, I believe that in multiple player situations it is good to tend as much as possible towards cooperation, because that strategy is most visibly the strategy of ‘the good example’.

Let’s look at some political animal rights issues to discuss the importance of this universalist imperative.

7.2.1 The argument of futility

A lot of meat eating people object that if they became vegetarians or vegans, the impact on the food market and the livestock sector will be negligible: not a single cow will be spared. However, the rule says that if everyone became vegan, then the end situation will be one without a livestock industry, which is better according to the three principles of equality. So therefore any individual has a moral duty to give the good example and become vegan.

7.2.2 Tit-for-what?

Suppose someone kills and eats ten small animals (chickens), unless I kill a big animal (a cow) and give it to him. Minimizing violations of basic rights and well-being implies that it is better to use one big animal as merely a means, than to use ten small animals as merely a means. So I should kill a big animal and give the meat to that person? This becomes a subtle issue. What if I take the conditional rule: “Do not kill animals, unless others kill small animals for consumption and you can reduce their killing by killing a big animal yourself and sell the animal’s products to those people”? Universalizing this rule in an ideal world will generate the best outcome, because no animal will be killed. The unconditional rule “Do not kill animals” also generates the best outcome in an ideal world. But in a non-ideal world, the conditional rule will be better.⁹

The argument to kill a big animal is similar to an argument given by domestic fur farmers: “If we don’t produce fur, then people will buy fur from countries with weaker animal welfare laws. I can produce cheaper and animal-friendlier fur that will

occasionally cooperating, even when the other player defects. If the other player plays tit-for-tat (with forgiving) as well, both players can escape the circle of defection.

⁹ Williams (Smart & Williams, 1973, p.97) offered a similar thought experiment to counter utilitarianism: should George accept a job at a laboratory for chemical warfare if refusing the job implies that another person takes the job and will do the unethical research with far greater zeal?

outcompete the fur from those horrible foreign fur farms. As a consequence, those fur farms have to lower their production. Hence, my production of fur will decrease the total animal suffering in the world.” Of course there will be other political strategies to decrease animal suffering (e.g. negotiations, import restrictions), but for the sake of the argument, suppose that those domestic farmers are right. Those domestic fur farmers work in a non-ideal world (non-compliance of foreign fur farms). Hence, they could use the conditional rule not to produce fur, unless it outcompetes worse fur production.

However, I do not believe that the conditional rule should be followed. There are two replies to this, one from a deontological (mere means) perspective, the other from a game theoretic perspective. The deontological consideration goes as follows: If I kill that big animal, then the animal will be used as merely a means in *two* ways: as consumption product by the other person (the other person uses the meat of the animal) and as ransom or a medium of exchange by me (I use the animal in order to save the lives of the other small animals). We can say that this double use, and especially the new use as medium of exchange, is never permitted, not even in non-ideal situations.

Related to this is the issue of the (non-)consequentialism of deontic principles: do we have a duty to minimize basic rights violations by violating someone’s basic right? We can say that if we violate the basic right of the big animal in order to stop ten basic rights violations of the ten small animals, the former basic right violation (of the big animal) counts heavier than a latter basic right violation (of a smaller animal) just as a latter basic right violation counts heavier than a right to live.¹⁰

As a second reply, the next table presents the outcomes from a game-theoretic point of view. The values represent the number of animals that stay alive. If both players cooperate, the eleven animals (the one big and ten small animals) will live. But player 2 wants to kill ten small animals. If player 1 refuses to sacrifice a big animal (i.e. if s/he cooperates), then ten small animals will die and the big animal lives. If player 1 kills a big animal, perhaps player 2 no longer kills the ten small animals. If both players ‘defect’, then all eleven animals will die.

		Player 2	
		Cooperate	Defect
Player 1	Cooperate	11	1
	Defect	10	0

¹⁰ In a previous section (6.5.1), we saw a mathematical equation with a basic rights term $-\sum r_i^X$. We can say that $r_i^X \gg r_j^Y$ when individual i (e.g. the big animal) is used in situation X as merely a means to stop the basic right violation of individual j (e.g. a small animal) that would have occurred in situation Y .

As with the above mentioned iterated prisoner's dilemma game, player 1 can play the strategy of tit-for-tat. But in contrast to the prisoner's dilemma, the new game has the worst outcome when both players defect. Hence, the retaliation strategy of player 1 is tricky. If player two does not cooperate, we end up with the worst outcome. But if player 2 cooperates (does not kill ten small animals), then tit-for-tat requires that player 1 also cooperates in the next round (does not kill a big animal).

The example of the fur farmer clearly demonstrates where this game strategy leads us. If the domestic fur farmer succeeds in out-competing the foreign fur production, then the domestic farmer has to stop his fur production. But then the foreign farmers might get new market space. The foreign production increases again, the foreigners defect, and the domestic fur farmer will defect again by producing fur, resulting in a decrease of the foreign fur production. This results in a continuous, high frequency cycle of quick changes between defection (domestic fur production) and cooperation (no domestic fur production).

There are two ways out of this cycle: (almost) always defect or (almost) always cooperate. The first way is a choice of continuous defection. Instead of a tit-for-tat strategy, it becomes a tit-for-what strategy. This strategy follows a rule: always defect if you *believe* the other person would defect when you cooperate. Player 1 might *believe* that player 2 will defect again and again once player 1 cooperates. In that case, player one might decide to continue defecting, no matter what player 2 does. The domestic fur farmer continues his fur production, even when the foreign production is out-competed.

However, this strategy of continuous defection has two problems. First, how can player 1 know that player 2 will cooperate even when player 1 would cooperate? The belief of player 1 that mutual cooperation is impossible, can never be disproven as long as player 1 keeps on defecting: player 1 does not even give player 2 the opportunity to demonstrate his unconditional cooperation. If player 2 would decide to cooperate even when player 1 cooperates, player 1 will never be able to know this if s/he always defects because of a false belief.

A second danger of such tit-for-what strategies is that outsiders cannot easily infer the rule or true motives of those domestic fur farmers: the rule depends on what player 1 believes about player 2, instead of what player 2 did. But outsiders cannot get reliable access to what player 1 believes. The fur farmers might lie when using the above argument: their true intentions might just be to sell fur, not to fight for animal rights. Related to this is the importance of giving the good example. I believe a lot of people would have difficulties in seeing the good example behind the strategy of trying to stop fur production by producing fur yourself. Therefore, in such fur farmers' games, I believe it is better to apply the rule to always cooperate.

In conclusion, in non-ideal situations we should 1) not introduce a new use as merely a means (e.g. a use as ransom), 2) not follow a tit-for-tat strategy if it result in a high

frequency cycle of cooperation and defection and 3) not follow a tit-for-what strategy (continuous defection) because of the risk of false beliefs and the lack of clarity of intentions behind the defection.

7.2.3 Prohibition laws

Consider a prohibition law: the government will punish anyone who eats, buys or sells meat. Imagine that if our government enforces this law, a black market of animal products will be generated. These products are smuggled illegally from a foreign country, where the rights of animals are violated on a much larger scale than the rights violations in our domestic livestock industry before the prohibition. In other words: imagine that animal products are subject to the same ‘iron law of prohibition’, like alcohol products. This iron law says that, in many countries, prohibiting alcohol production and trade will result in more alcohol abuse. Making alcohol illegal will generate worse consequences. Of course, there is a difference between illegal alcohol production and trade on the one hand, and illegal livestock production and trade on the other. Likely, the livestock prohibition will not be subject to the iron law of prohibition, because some basic conditions that led to the iron law of prohibition for alcohol are not met in the prohibition of animal products. A government will be able to forbid the trade and production of animal products, just as with human products. Due to characteristics of livestock industry, illegal livestock production and trade is much easier to find than illegal alcohol distillers and bootleggers. But for the sake of the argument: imagine that prohibition of animal products would make matters worse in terms of animal rights violations. What should we do then? Should we give in to a kind of blackmail if meat eaters say that prohibition will result in worse animal rights violations overall? If they say that making meat illegal results in a temptation to eat more meat and a worsening of conditions for the animals (e.g. smaller cages, no more government control)?

If no-one trades, produces and consumes animal products, no basic rights will be violated, so therefore I should not buy, sell, produce or consume animal products. This rule applies to the ideal world of universal compliance, and, as it gives the best results in this ideal world, we should stick to this rule in a non-ideal world as well. But what about prohibiting and punishing others who produce or consume animal products? What about a government policy to make meat illegal? In the ideal world, prohibition and non-prohibition would be equally good, because there will be no-one to be punished. So prohibition and non-prohibition would result in the same consequences in the ideal world. We have a tie between two rules: prohibition and non-prohibition. The best rule should now be derived by a reality check. The move to a non-ideal world will be a tie-breaker. What if, as in reality, not everyone will follow the rule to abstain from animal products? If it would occur that a prohibition in a non-ideal world would result in more

animal rights violations, compared to a non-prohibition, then non-prohibition should be preferred. In other words: if animal products would be subject to the iron law of prohibition (as with alcohol), if prohibiting and punishing the production and trade in animal products would result in worse animal rights violations (which is likely not the case), then prohibition and punishment would not be a government's duty. In this case, our only duty will be to abstain from buying and selling animal products ourselves, but we should not prohibit and punish others.

7.2.4 Self-defense against culpable attackers and innocent threats

Non-ideal theory also deals with situations of self-defense against attackers who violate a moral rule. To discuss this issue, let us consider a trolley problem that represents a rather broad picture of self-defense against both culpable and innocent threats.

Imagine some people start driving a trolley, and this trolley is then heading towards a number of potential victims on the main track. If the people in the trolley have the intention to hit the victims, and if there is no sufficiently strong justification to hit the victims, then those people in the trolley are culpable attackers. In general, a culpable attacker is someone who consciously wants to do an action that violates a moral rule; in particular an action that decreases the welfare function or the more general moral weight (that contains the welfare function plus some additional terms that represent violations of the mere means principle). It might also be the case that the people in the trolley are unaware of the victims on the main track (i.e. they are misinformed), are coerced to start the trolley (i.e. they act under duress), or are innocent in some other way (i.e. they are insane or hypnotized). In those cases, the people in the trolley are innocent threats.

Furthermore, imagine that the potential victims can save themselves by turning a switch that sends the trolley to a side track. This side track ends in a ravine, so turning the switch will result in harming the people sitting in the trolley. But on this side track, there may be a number of innocent bystanders who also might be harmed when the switch is turned.

From a moral point of view, I make no distinction between innocent bystanders on the side track and innocent threats in the trolley. So in general we have a number of N_v potential victims on the main track, N_c culpable attackers in the trolley and N_i innocent people in the trolley and on the side track. The level of damage per person that might

befall each of the potential victims if they don't defend themselves is d_v .¹¹ Hence, the total damage of all the victims when they don't defend themselves is $N_v d_v$. Similarly, if the victims defend themselves by turning the switch, they cause total damage $N_i d_i^*$ and $N_c d_c^*$ to respectively the innocent people and the culpable attackers. (The * refers to the situation where the victims act in order to defend themselves.) To make it more general, we can also consider a damage $N_v d_v^*$ that the victims might still receive even when they defend themselves (for example when turning the switch would be harmful for the defending victims as well), and a damage $N_i d_i$ that the innocent people get when the victims do not defend themselves (for example when the innocent people on the side track are blown away by the trolley passing by at full speed on the close-by main track, or when the innocent people in the trolley get hurt when the trolley hits the victims on the main track).

The question now is: when is it allowed for the potential victims to defend themselves, respecting a proportionality condition on self-defense? There are three proportionality constraints that the defending victims should respect. First, the most obvious constraint: in their defense, the potential victims should take the option that avoids any unnecessary harm. If the same results could be achieved with a lesser harm, then they should opt for the defensive action that causes the lesser harm.

Second, note that the culpable attackers, by consciously wanting to violate a moral rule (i.e. consciously wanting to decrease the welfare function or moral weight), place themselves in a sense outside of morality. In that case, the defending victims should avoid decreasing the restricted welfare function or moral weight. This restriction means that the harms (the loss of lifetime well-being) suffered by the culpable attackers as a result of the defensive action of the potential victims, are not included in the equation. In other words, in the restricted welfare function the levels of lifetime well-being of the culpable attackers is constant and equal to the levels they would have when they did not violate the moral rule. Hence, a defensive action by the potential victims is allowed if the following inequality constraint is satisfied:

$$N_i d_i^* + N_v d_v^* \leq N_i d_i + N_v d_v.$$

This means that the total harm of the defensive action (excluding the culpable attackers) should be lower than the harm suffered by the victims and innocent people when the victims don't defend themselves. Here, again, the harm is measured by the

¹¹ Looking at the welfare function, this damage per person might be written as $d_v = f(\Delta x_v) = (\Delta x_v)^p$, with Δx_v the loss of lifetime well-being of a victim. When a victim is used as merely a means, the damage becomes much bigger, because d_v includes the parameter r_v that measures the victim's basic right violation.

decrease in welfare function or moral weight. The harms suffered by the culpable attackers are not included in this inequality constraint.¹²

What if we suppose that the culpable attackers should be treated exactly as innocent threats? Imagine there are five culpable attackers ($N_C=5$) and only one potential victim ($N_V=1$). The lifetime well-being of the five attackers might trump the lifetime well-being of the one potential victim. This would mean that if the life of the victim is at stake, the defensive action of the victim should not result in the death of more than one of the attackers. This seems counter-intuitive. If you are attacked by five killers, you are allowed to kill all of them in self-defense, if killing them is the least harmful option you have in your defense.

As an example of animal ethics, consider a person being attacked by five predators. If the only option of self-defense is to kill everyone of those five predators, then my moral intuition says that the attacked person is allowed to kill all five predators, even when the death of five predators might be worse (in terms of loss of lifetime well-being) than the death of one prey.¹³

That is why the harms caused to the culpable attackers should not be included in the welfare function or the moral weight. But the culpable attackers are not completely placed outside of morality. True, the culpable attackers consciously violate a moral rule against harming others. The total damage they cause is $D=N_V d_V + N_I d_I$. Are the victims in their defense allowed to do anything with the culpable attackers? No, there is a third proportionality constraint that they should respect, given by the following inequality

$$d_C^* \leq D = N_V d_V + N_I d_I.$$

In other words: to make the permissibility of self-defense in line with our moral intuitions, we can state that each single culpable attacker who consciously wants to cause a total level of damage D is liable to that amount of damage in the defensive action

¹² The permissibility of self-defense is determined by the choice between the welfare function $W(x_b, x_v, x_c)$ where the potential victims (who have a well-being x_v) do not defend themselves, and $W^*(x_b^*, x_v^*, x_c)$ where the potential victims defend themselves. Note that the culpable attackers have the constant lifetime well-being x_c^* instead of x_c and x_c^* . This constant lifetime well-being is the level the culpable attackers would have when they did not attack, i.e. when they did not violate a moral rule. Similarly, in the welfare function of the psychological connectedness description, we fix the value of $\hat{\mu}'_{C(t)}$, i.e. the integrated well-being of a person who has at time t the culpable intention to violate a moral rule, when that person did not violate that rule (as indicated by the quotation mark). The extra constraint on self-defense in this formulation becomes more complicated. One option is for example: $d_{C(t)}^* = \hat{\mu}'_{C(t)} - \hat{\mu}_{C(t)}^* \leq \max_t \sum_{\pi(t)}^{N_V(t)} (\hat{\mu}'_{\pi(t)} - \hat{\mu}_{\pi(t)}^*)$, i.e. the damage to the culpable person (the difference of the integrated well-being between the no-attack situation and the defense situation) is less than the time-maximum of the sum of damages of all victims (the difference between the no-attack situation and the no-defense situation).

¹³ In chapter Chapter 10 I will argue that predators are still allowed to hunt. This does not contradict the permissibility of prey to defend themselves.

by the victims. Hence, the maximum permissible level of damage d_c^* that the victims can cause to each one of the culpable attackers is D .

This third proportionality constraint does not follow from the maximization of the welfare function or the moral weight. It is a constraint that is applicable only to non-ideal situations, where some people consciously want to violate the welfare function. So we should follow the following conditional rule: “Maximize the welfare function, unless others (the culpable attackers) consciously don’t. In that case, maximize the restricted welfare function (excluding the harms suffered by the culpable attackers) and respect the proportionality constraint on the harms of the culpable attackers.” If everyone complies with this rule (the ideal situation), the welfare function is obviously maximized. But compared to the unconditional rule “Maximize the welfare function”, this conditional rule works better in non-ideal situations: it better fits with our moral intuitions about self-defense.

7.2.5 Summary

In summary, we start by looking at an ideal theory. In a Utopian world with universal compliance, we could derive the following rule that generates the best outcome relative to our theories of equality: don’t consume, trade or produce animal products. This rule should also be applicable to a non-ideal world. On the other hand, ideal theory could not make a decision between the rules of prohibition versus non-prohibition. Then we have to look at a non-ideal world lacking universal compliance. In such a world, under certain specific conditions (resulting in an iron law of prohibition), non-prohibition (non-punishment) would be preferable. These conditions might be met for alcohol production and trade, but likely not for animal products. Non-ideal theory is also relevant to derive proportionality constraints in defensive action.

So far for universalization with respect to the agents. In the next section, I will shed more light on the second kind of universalization, with respect to the patients. This relates to discrimination as a violation of the formal principle of equality. But discrimination is also related to ideologies of hierarchic dualisms. Before I argue in the next chapter that speciesism is a discriminatory hierarchic dualism, I will briefly discuss the properties of hierarchic dualisms.

7.3 Formal equality, discrimination and hierarchic dualism

Formal equality is related to the notion of discrimination. Let's define discrimination as causing a disadvantage to an individual (or a group), based on a value-laden distinction between individuals (or groups), where the distinction is not justified or refers to properties of the individuals (or groups) that are not deemed morally relevant in that situation.

In other words: a person A discriminates B against C, if A believes (and acts on the belief) that B has lower value than C (meaning C *should* have more rights, advantages or opportunities than B), where this value difference has no justification or is derived from properties of B and C that are not morally relevant or are not an acceptable motive for the decisions and behavior of A.

The question is: what are morally (ir)relevant properties or criteria? In part 3 I will argue that being human is not a relevant property, and sentience is. But for now, if we place it in the QMM-framework, the answer is simple: morally relevant properties are all properties that are related to improving the value of life of all individuals who have a well-being, in line with the QMM-principle. We argued that desert based and resource based principles follow from QMM. So morally relevant properties are amongst other things: desert (contribution to the value of life of others), effort, incurred costs and personal responsibility.

We now move from the QMM-principle to the principle of tolerated choice. In the burning house, we would save our own child instead of an unknown child. Now, the tricky point is that your child is not more deserving or responsible, simply because it is your child. Actually, the fact that it is your child is not important in the light of QMM-theory. Of course, if you lose your child, your value of life will be affected. But don't forget that the parents of the other child in the burning house will also feel sad when their child dies. The death of your child is as bad as the death of the other child, if we look at QMM-theory.

So do we have discrimination? In some sense yes: there is an emotional inequality in our behavior towards different children. Yet, there can still be some subtle form of equality present, which is the tolerated choice equality. It is related to the words 'value-laden' in the definition of discrimination. What do we mean by this? Suppose the parent of the other child passes the burning house, and saves his child. There are two ways how you can react. You can say that what that person did was immoral, because your child has more intrinsic value, a higher moral status or a stronger moral right to live. Or you can say that, although you regret that your child died, you accept and tolerate the choice of that person to save his child. In the latter case, you and the other parent are in some sense equal, and therefore your children inherit a tolerated choice equality, although there is an emotional inequality from your point of view.

As we have seen, there is one subtlety with tolerated choice equality: what if a white employer refuses to give work to a black person? If the judgment of the employer is based on prejudice, the employer makes a value-laden distinction between white and black employees, which is racist discrimination.

Hence, tolerated choice equality should be distinguished from a moral value-laden inequality, which results from a discriminatory ideology such as racism, sexism or specicism (Ryder, 1975). These kind of ideologies are hierarchical dualisms (Plumwood, 1993) between an upper side (the oppressors) and a lower side (the oppressed). Hierarchical dualisms can be characterized by one or more of the following properties (this is a small extension of the theory of Plumwood).

1) The lower side is radically excluded from the upper side, by believing that there is a deep gap between the two sides. Any overlapping between the two sides is denied.

2) The lower side is negatively defined: the oppressed lack the properties which are used by the oppressors to justify the oppression.

3) The lower side is homogenized, individual differences between people from the lower side are denied, by use of e.g. stereotyping.

4) The lower side is marginalized: the oppressors do not show care and empathy. The personalities and needs of the oppressed are denied or scorned.

5) The lower side is unjustly criminalized, they are the scapegoats. The innocence of the lower side is denied.

6) The lower side is instrumentalized (objectified), they are used as tools, as means to the ends of the upper side. The intrinsic value of the lower side is denied.

The first three characteristics are psychological mechanisms to sustain and justify unequal treatment. These mechanisms result in violations of the tolerated choice principle. Characteristics 4 and 5 are violations of the QMM-principle. The sixth is a violation of the basic right principle. So we see that our three material principles of equality are related to ideologies of hierarchical dualisms. When one or more of these characteristics are present, there is a value-laden difference, and we can speak of immoral discrimination.

In the next part of this dissertation, we move from normative ethics to applied ethics, in particular to applied animal ethics. An animal ethic gives a fundamental critique on the ideology of speciesism. Looking at the above six characteristics of hierarchical dualisms, all of them are present in our current speciesist society, just like they are present in racist and sexist societies. Although this is not yet proof that speciesism is a kind of immoral discrimination comparable to racism and sexism, it strongly enforces that idea. The proof will be given in the next chapter.

Part 3 Animal ethics

Chapter 8 Speciesism as a moral illusion¹

8.1 The current situation: patho-anthropocentrism

In this chapter I argue that speciesism is based on a moral illusion. The species boundary is not morally relevant. Sentience is the morally relevant criterion. Our current speciesist society is discriminating non-human sentient beings (animals).

Anthropocentrism is the ideology that ascribes a central moral value to all humans (individuals belonging to the species *Homo sapiens*). However, not all entities with human DNA have an equally high moral status. In a lot of countries, abortion is legal² (not murder), and human embryos are also sometimes used in stem cell research and therapy. An often heard justification for this use of embryos (as merely a means, because those embryos die), is that those individuals have not yet developed a complex central nervous system that gives them the capacity to feel and be conscious. Apart from conservative religious people, most people are in favor of legal abortion and embryonic stem cell research and therapy. As a lot of other people, I do not consider a fertilized human egg cell as a human being. True, it has the complete genome of a human being, but by that criterion a skin cell would also be a human being. At most, the fertilized egg cell can (under the right circumstances) develop into a human being.

It appears that there is some fuzziness when exactly we call a being a human being. A lot of people are in favor of using non-sentient embryonic humanlike beings as merely a means in scientific research to help other people. So, the criterion of sentience already drips in the ideology of anthropocentrism. Feelings are relevant as well, so we might

¹ Sections 8.2, 8.3, 8.4, 8.5, 8.6 and 8.7 are based on *Can Speciesism be a Moral Illusion?*, submitted to Journal of Moral Philosophy. Section 8.8 is based on Bruers (2013), *Speciesism as a Moral Heuristic*, *Philosophia*.

² However, the legality of abortion (especially in cases of rape) does not yet imply that human fetuses and embryos have a low moral status. One could argue that the fetus is not allowed to use the pregnant woman as merely a means. The pregnant woman has autonomy over her body.

rather call the current ideology ‘patho-anthropocentrism’ (although this still might be an oversimplification of our current society). This states that sentient humans have the basic right and belong to the moral community.³ **Figure 12** illustrates this.

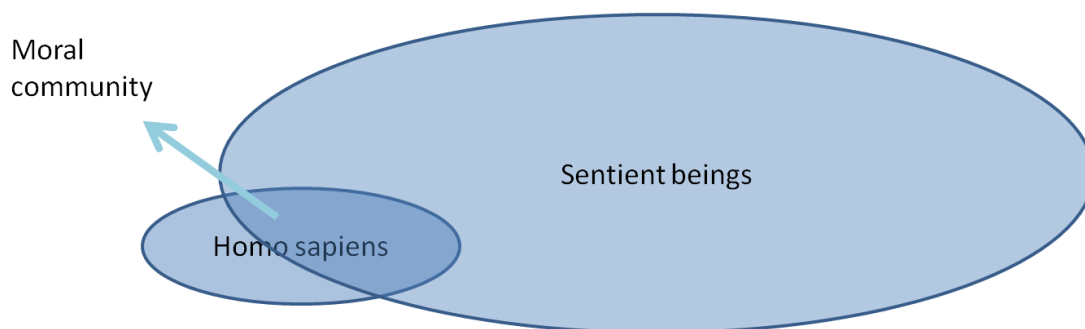


Figure 12: the moral community according to patho-anthropocentrism

I am now going to argue that it is a kind of discrimination to include the criterion ‘*Homo sapiens*’. If we value a sentient *Homo sapiens* higher than another sentient being, then this is a kind of moral illusion.

8.2 Moral illusions and discrimination

The Müller-Lyer optical illusion, presented in the first part of this dissertation, is a very illuminating analogy of the moral illusion of discrimination. It can be used to represent speciesism (Ryder, 1975; Singer, 1975; Regan, 1983). The suggested correspondence can be seen as follows.

The two horizontal lines in the figure can be interpreted as the respective moral statuses (or intrinsic values) of a non-human animal and a human. The length of the line segment is the analogue of the level of moral status. A lot of people have the intuition that the moral status of a pig is lower than the moral status of a human, just like a lot of people have the intuition that the upper line segment is shorter than the lower one.

The small lines (arrowheads) represent the morally irrelevant properties, such as specific genes, physical appearance (such as having a tail), or having the capacity to get

³ Note that also conservative religious people value sentience in some sense, because sentience is an important aspect of their notion of the soul, and they believe a soul enters the body at the moment of conception. Scientifically, this is not true: there is no evidence for this and it is against a coherent scientific theory that says that brains generate sentience.

fertile offspring with someone. Those things do not matter, just like having a white skin color or having a penis doesn't matter for moral status. But according to racists, skin color does influence their intuitive judgment that black people have a lower moral status. Just like those morally irrelevant criteria, the arrowheads in the figure are geometrically irrelevant as well. They do not determine the lengths of the two horizontal line segments. As we have seen, this is the notion of context independence.

We can now make the analogy with discrimination: If you judge that the two horizontal lines in the Müller-Lyer illusion are dissimilar whereas in reality they are not, then you discriminate. If the lines are in reality dissimilar, then judging them to be different in length is not discrimination.

We have seen that not everyone is susceptible to the Müller-Lyer illusion (some indigenous people have no differential judgment when they grow up in environments without straight lines of tables and staircases), and we have seen how the underlying psychological mechanism works (3D adaptation of a 2D image). We also know something about the psychological mechanisms behind discrimination. It is based on an in-group-out-group bias (Tajfel, 1981; Whitley & Kite, 2010). Although in-group-out-group value judgments occur intuitively, several studies (Kurzban et al., 2001; Cosmides et al., 2003) demonstrated that the choice of in-group-out-group (e.g. based on race) is not inborn, but is culturally dependent and can be influenced by changing cultures. Speciesism is also culturally dependant. In some cultures (e.g. Jainism) and in a big part of the animal rights movement, people do not (or no longer) have the prejudicial judgment that the moral status of humans is higher than other animals. The intuitive judgment is not universal and not inborn. But people growing up in a speciesist society assimilate this ideology until they get this discriminating moral intuition. The same happened with people growing up in racist societies. They often perceive their in-group-out-group distinction as being natural, but it is not.

Similarly, people growing up in an environment with houses and tables, often see straight edges, and therefore they assimilate optical intuitions about lengths of lines. The *disposition* for such an assimilation process is natural (inborn), but the *result* is not. In-group-out-group thinking is natural, inborn and universal as well, but the result (which group is the in-group), is not.

This means that ideologies such as white-dominant racism, male-dominant sexism or human-dominant speciesism are strongly culturally determined. These ideologies are not universal, and perhaps the underlying intuitions behind those ideologies are more flexible and can change more rapidly than some of our 'deeper' moral instincts. The dividing line between the ingroup and the outgroup can be influenced by society and is vulnerable to change. It remains to be seen whether intuitions behind e.g. the mere means principle are equally flexible and influenced by culture. Although there remains some experimental controversy (Sachdeva et al., 2011), I slightly expect that those intuitions behind e.g. the mere means and QMM principles are more universally 'hard

wired' in our brains. Some evidence for this 'universal moral grammar' hypothesis can be found in Mikhail (2000; 2007), Hauser et al. (2008) and O'Neill & Petrinovich (1998).

8.3 How do we know whether speciesism is a moral illusion?

In the Muller-Lyer illusion we had reliable instruments to demonstrate that it is an illusion: we could use a measuring stick or something to cover the small lines. In ethics, our reliable instruments are valid arguments, so we need arguments based on ethical principles that form a coherent system and are compatible with our strongest moral intuitions. These two requirements, coherence and compatibility, are very important. In the optical illusion, we used a coherent theory of geometry, which is compatible with two very strong intuitions: translation invariance and context independence. If we want to argue that speciesism is really discrimination, we need tools of similar power.

And we have those tools. We will present no less than ten arguments: five arguments against the species boundary (to demonstrate that the criterion *Homo sapiens* is not morally relevant), and five other arguments to show that sentience is really important. Those ten arguments are also based on moral intuitions, some of them quite strong. And they form a coherent theory: the arguments mutually support each other. For the speciesist it would be very difficult to attack this system.

The five arguments against the species boundary can be compared with the principle of context independence in the Müller-Lyer illusion. The five other arguments in favor of sentience can be compared with the principle of translation invariance. So, as in the Müller-Lyer figure, we have one intuition (the human-animal value difference in ethics or the length difference in geometry) which is in contradiction with several other intuitions (e.g. the importance of impartiality in ethics and the context independence in geometry). And as in the optical illusion, in the moral illusion we have two options. First, we could abandon all ten arguments and their underlying strong moral intuitions. This would save our intuitive judgment about the human-animal value distinction. Or, second, we could admit that this value distinction is an illusion, and we can save the stronger moral intuitions. I believe that the combination of the latter ten intuitions (underlying the ten arguments) is stronger than the one intuition about the value of humans versus animals. So the easiest thing to do is to acknowledge that this human-animal value distinction is a moral illusion, similar to the optical illusion in geometry. This acknowledgment is furthermore acceptable if we keep in mind that the in-group-out-group distinction is – just as the Müller-Lyer illusion – not inborn but culturally dependent. And the fact that even after realizing it is an illusion, the intuition persists, does not justify this intuition. The Müller-Lyer intuition, too, was 'cognitively

impenetrable' (Pylyshyn, 1999): our intuition keeps on saying the one line appears to be longer than the other, even after we have measured them.

Before we give the ten arguments, I suggest we first have a look at the following figure.

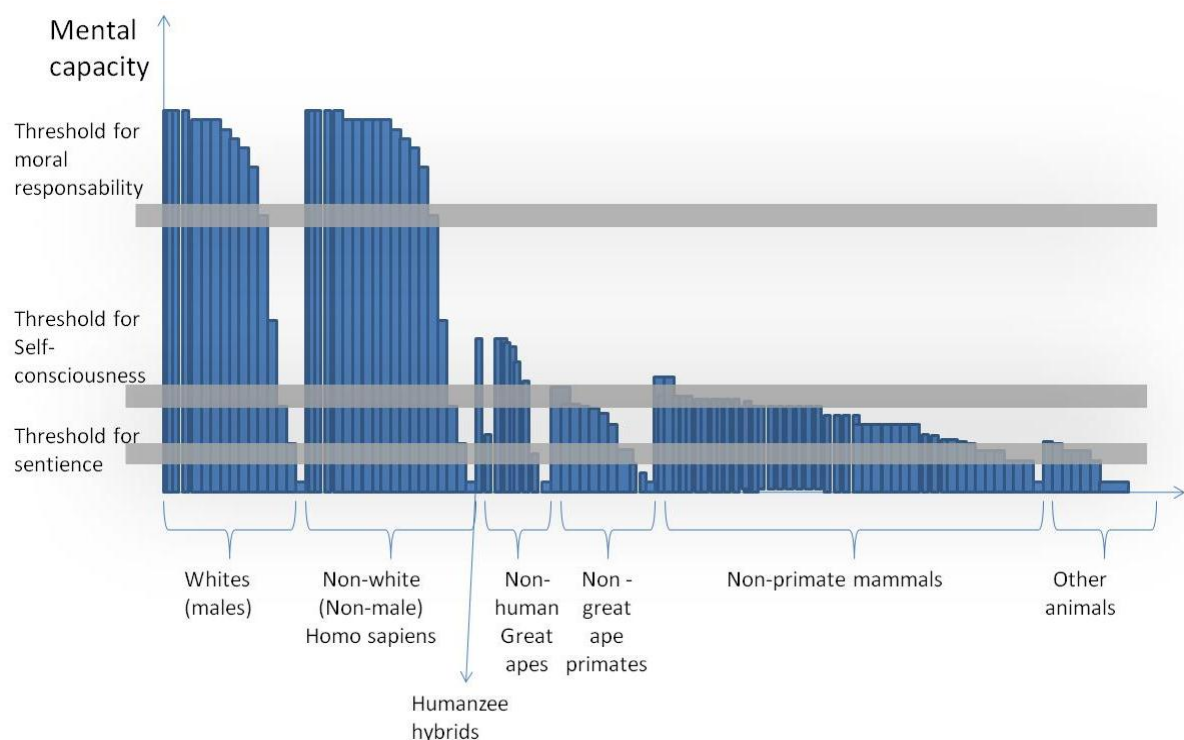


Figure 13: mental capacities of responsive beings

Each vertical bar in the above figure represents a responsive being (an animal with a nervous system). We can (for simplicity's sake) consider three mental capacities: sentience, self-consciousness and moral consciousness. These capacities are represented as grey horizontal bars, because the thresholds are often vague. When does a being exactly have a moral consciousness? Although the threshold is vague, we can clearly see that only humans (but not all of them!) possess a moral consciousness. Non-human great apes (but not all of them!) possess self-consciousness. And quite a lot of animals possess sentience.

Looking at this figure, how do we decide the moral community? Which vertical bars (individuals) do we take into account? Racists only considered white people (and perhaps not all white people). Speciesists could consider all *Homo sapiens*. But what about great apes and potential human-animal hybrids? (See next section to learn more about such hybrids.)

Another possibility is to take only the vertical bars that reach above some threshold. E.g. Kant (1785), Rawls (1971), Cohen (1997), Scruton (1998) and many other anti-animal rights philosophers took only those beings who have a moral consciousness. So they did not know what to do with the humans below the threshold of moral agency. A lot of

anti-animal rights philosophers refer to such mental capacities that mentally disabled persons lack. For people who are concerned about the rights of those disabled persons, this approach is often offensive.

But I suppose that the argument from marginal cases is valid (Dombrowski, 1997), meaning that mentally disabled human orphans have a high moral status and basic rights (as they have in modern societies). In other words: criteria such as higher mental capacities, language, self-consciousness, moral consciousness or social bonds are already excluded and are morally irrelevant in the contexts I am thinking about, i.e. when it comes to our treatment of animals for food or experiments.

The great ape project (Singer, 1993) might take self-consciousness as the relevant threshold. But even proponents of this project exclude some seriously mentally disabled humans. I will argue that we'd better take the threshold of sentience. This includes almost all mentally disabled humans. Only those human beings who are merely alive or responsive but not sentient, would be excluded. (But as we have seen in the chapter about the basic right, also responsive and living beings should be given some weaker version of the basic right.)

In the next section I give five arguments against the species boundary, and five other arguments in favor of the sentience criterion. Some important remarks are in order before I present the arguments.

1) With moral status, I am referring to an agent-independent moral status, which means agent-dependent relational aspects are not (and cannot be) the basis of the human-animal distinction in our current society. Consider the burning-house dilemma: I have to choose between saving my child or the dog. I prefer to save my child, because I feel a stronger connection or relationship with my child than with the dog. I also feel a stronger connection with my child than with yours, so again I would prefer to save my child. Some partiality might be allowed, as long as we tolerate similar levels of partiality of other moral agents. I would tolerate your choice to save your child instead of mine. But if I believe that my child has a higher moral status independent from my relationship with my child, I would not tolerate your choice to save someone with a lower moral status, and this will become a kind of discrimination.

The agent-relative relational moral status is important in an ethic of care, but cannot explain the huge gap in moral status between humans and non-human animals in our current society. The partiality reflected in this gap is intolerably big, because we would not tolerate similar partiality in the other direction, where animals would have the status that humans have now, and vice versa. We would not tolerate that a non-human animal would kill and eat a human just for taste. Similarly, I would tolerate your choice to save your child instead of mine from the burning house, but I would definitely not tolerate that you kill my child for the gustatory pleasures of your child. Your special relation with your child does not allow you to do the latter.

2) I am supposing a moral individualism as defined by McMahan (2005) and Rachels (1990): the agent-independent moral status of an individual is uniquely determined by his/her own particular characteristics. Hence, his/her group memberships are irrelevant. In fact, this assumption is an analogue of the principle of context independence in the Müller-Lyer figure: the length of a line segment is uniquely determined by an intrinsic property and its belonging to the group of line segments with outward pointing arrowheads should have no influence. Moral individualism takes an eraser to gum out the irrelevant elements such as group memberships.

3) When I refer to species, I refer to the biological (scientific) notion of a species, and when I refer to humans, I refer to the species *Homo sapiens*. This might seem obvious, but I don't want to target a straw man: people who defend speciesism refer to 'humans' and 'species', and I don't know of any other notion that makes sense of these terms apart from the biological notion. Some philosophers defend speciesism, not by relying on the scientific notion of species, but rather by using a 'folk' notion of 'human beings' as a (natural) kind (e.g. Chappell, 2011). However, those philosophers are not clear on whether the group of human beings (according to the folk notion) equals the group of *Homo sapiens* (according to the biological notion). If those two groups are identical, there are two options: either this equality is a mere coincidence, or there is a (causal) explanation of why these two groups coincide. The former case is very unlikely, the latter case means that the folk notion can be reduced to the scientific notion (or perhaps the scientific notion is based on the folk notion, which means that biologists would be guided by folk intuitions). If the two groups are not identical, those philosophers should clarify what which human beings are not *Homo sapiens* or which *Homo sapiens* are not human beings.

In the section on essentialism and heuristics, I will elaborate more on folk notions of natural kinds.

4) I do not assume that all speciesists (people believing in the status gap between humans and animals) believe that species is the only morally relevant criterion for moral status. In fact, speciesists who defend stem cell research or abortion likely include sentience as necessary criterion, next to species membership. According to them, non-sentient humans such as fertilized embryos have a lower moral status. And most speciesists also support animal welfare laws, which implies again that sentience has some importance.

8.4 Five arguments against the species boundary

The arguments that I will present, are based on a moral intuition: in order to avoid the risk of opportunism in our ethics, we should try to avoid adding arbitrary, artificial, farfetched or fuzzy elements to our ethical system. The arguments are strongly based on the biological sciences. As Rachels (1990) and Hull (1986) demonstrated, especially the Darwinian paradigm undermines some metaphysical beliefs about 'humanity', 'human nature' or 'the human kind'.

1) The biological species boundary is arbitrary. There are two kinds of arbitrariness: a vertical and a horizontal one. The vertical arbitrariness asks the question why we should select 'species' from the list of biological categories? I belong to the kingdom of animals, the phylum of chordates and vertebrates, the class of mammals, the infraclass of eutheria, the order of primates, the suborder of dry-nosed primates, the infraorder of simians, the superfamily of Hominoidea, the family of great apes, the genus *Homo*, the species *Homo sapiens*, the subspecies *Homo sapiens sapiens* and the ethnic group of whites. There are different genetic affinities. It is arbitrary to pick out the species. One could reply that individuals within a species are characterized by similarities (in terms of e.g. common genetic heritage, physiology or behavior), but the same goes for the other biological categories. There are multiple similarities, and they come in degrees. It remains arbitrary to pick out some similarities as being important, and also to pick out a specific degree of those similarities as being important.

Next to vertical arbitrariness, horizontal arbitrariness asks that, if we select species as the relevant category, why should we take one species instead of another? Also the choice for a specific species amongst the many species is arbitrary.

Note again that the existence of mentally disabled humans demonstrates that there is no exact correlation between mental capacities and species. Note also that referring to a *folk* concept of 'human being' as a *kind* (as proposed by e.g. Chappell, 2011) instead of a specific biological conception currently favored by science, would not avoid the arbitrariness. Not only is the notion of a 'kind' ambiguous (is it a natural property or a construct?), each clarification of this notion results in the observation that there are many possible 'kinds'. If 'human being' is a kind, then for example 'primate', 'two-legged being' or 'caucasian' can be kinds as well. It is not clear why the kind of human beings is an exceptional kind, compared to the many other possible kinds that one can think of.

Also other discriminations have a double arbitrariness. For example in religion, there are the groups of Benedictines, Roman Catholics, Catholics, Christians, Abrahamists,... Why should people in Northern-Ireland pick the third category in this row? And within this category, why should they pick Catholics instead of Protestants? The same applies to ethnicity: it is arbitrary to pick the second category in the ranking of a) people from

Antwerp, b) the Flemish people, c) Belgians and d) Europeans. And it is arbitrary to prefer the Flemish people to the Walloon people in Belgium.

2) The biological definition of species is very complicated and too artificial and farfetched to be used in a moral system. One of the many definitions of species refers to the possibility of interbreeding and getting fertile offspring. But why should this possibility be relevant? It is too farfetched to say that a being has a moral status if its close relatives (parents) could have gotten fertile offspring with some other morally relevant beings. (I refer to its close relatives because the individual itself could be infertile.) It is unfair that an individual gets rights because his parents are able to do something with others. It is unjust to take a principle where non-human animals simply have bad luck having the wrong parents.

Related to this is the issue of so called ring species such as the *Larus* gulls, the *Ensatina* salamanders or the Greenish Warbler. Such ring species consist of different populations, whereby A can get fertile offspring with B, B with C, C with D, but D not anymore with A. Just as populations of ring species are spatially related to each other, we can say that all species in nature are temporally related in a similar way. Ring species "are only showing us in the spatial dimension something that must always happen in the time dimension." (Dawkins, 2004, p.303). Look at the phylogenetic tree. A modern *Homo sapiens* could have fertile offspring with an ancestor, that ancestor with an older ancestor, and so moving up a branch of the phylogenetic tree until we reach a common ancestor of both *Homo sapiens* and another species. Then we move down the branch of that other species. So there is a chain of populations connecting our species to any other species. The clue is that the higher moral status of A (a *Homo sapiens*) compared to D (an individual of another species) strongly depends on the fact that B and C are dead. Formulated this way, it becomes clear that such dependency on the accidental non-existence of individuals cannot be morally relevant.

Again one might object that speciesism does not refer to the biological notion of species, but to the folk concept of 'human being' as a kind. But this escape maneuver does not work either: the folk concept is perhaps even more complicated than the biological concept of a species. The folk concept of 'human being' might be based on a pattern recognition: when we are confronted with an object, we can spontaneously recognize the pattern and see a human being. But if such a pattern recognition is based on an algorithm, it is a highly complicated algorithm. The folk notion looks trivial, but it is not.

3) There is a potential fuzzy boundary: it is not unlikely that a human-chimpanzee hybrid (humanzee or chuman) can be born. Nearly 10% of mammal species can form interspecies hybrids. In the wild and in zoos, there exist lion-leopards, lion-tigers, camel-lamas, dolphin-killer whales, sheep-goats, grizzly bear-polar bears and off course the well known horse-donkeys (mules). If these are possible, and if the genetic distance between humans and chimpanzees is not larger than the distance between those

interbreeding species, it is possible that humanzees can be born. What would the moral status of this hybrid human be? There is an arbitrariness here as well. And what if a neanderthal (*Homo neanderthalensis*) would still exist? Would we give him the basic right? And what about other ancestors such as the *Australopithecus* and the *Homo habilis*? And there is more arbitrariness when we look at the possibility of human-animal chimeras. A chimera is an individual composed of genetically distinct cells that originate from human and animal zygotes. The body cells of chimeras can range from 100% human to 100% non-human. Where to draw the line of humanity? What would the moral status of such chimeric individuals be? And we could also genetically modify humans and animals. All of this blurs the line between humans and non-human animals. Science will not be able to propose criteria to determine whether beings such as neanderthals, hybrids, chimeras and genetically manipulated beings should be called 'human', just as scientists are not able to determine whether grains of sand should be called a heap. This fuzziness is a philosophical issue, not a scientific one.

Similar fuzzy boundaries occur in other kinds of discriminations: sexism is faced with different kinds of transsexuals and intersexuals (who have genital ambiguity or mixed chromosomal genotypes) and racism is faced with different kinds of mixed races such as mulattoes. This also makes it very complicated to define sex and race.

4) Species boundary refers to genes or appearance, and these are not morally relevant, because racism and sexism are also often based on genes or appearances and antiracism/antisexism states that such a basis is not morally relevant. If we say that skin color (or the genes that generate skin color) is not morally relevant, we should apply this rule consistently (universally) and state that no reference to appearance or genes is morally relevant when it comes down to someone's basic rights, as long as we do not have an argument that some appearances or genes are exceptional. A racist should be able to explain why skin color is morally relevant but e.g. hair color isn't, and if he can't explain it, then he should treat skin color as hair color. Otherwise we open the door for opportunism. So we should universalize the rule that genes and appearance are morally irrelevant for *everyone* in *all* situations related to basic rights violations. Also, there is no 'interest gene' connected to all and only to humans; there is no gene that makes a being to have interests.⁴

5) Belonging to a certain species instead of another is not something that we can choose, it is not something we achieved, it is beyond our responsibility. Belonging to a certain species is also not related to subjective needs and preferences. Hence, we should not be rewarded for belonging to a species. We do not deserve special treatment by

⁴ Recently, Liao (2010) developed a 'genetic basis for moral agency', but that approach was criticized by Grau (2010). I would add that this genetic basis account is as farfetched as genetics is complex.

having some specific genes. Giving a higher moral status to beings who did not choose to be born that way is in violation of the merit principle. If we are to be rewarded, it is not merely because we are born in some way rather than another, but because we either have a certain responsibility for an action (for example we did an effort or we contributed to something valuable) or we have needs and are able to subjectively experience and prefer things (for example we have a well-being and feel our needs). Moral advantages (rights, resources, opportunities,...) should be given to someone who deserves it or someone who needs it. But someone's species is not related to merit (responsibility) nor need. On the other hand, as we will see below, sentience is related to having subjective needs and preferences, so sentience is a reason to give someone a moral advantage.

Note that the above five arguments are very similar to the principle of context independence of the Müller-Lyer optical illusion. In the Müller-Lyer illusion, the irrelevant context (the arrowheads) was characterized by arbitrariness, artificiality and fuzziness. As the species distinction has those same characteristics, we can say that the species are part of an irrelevant context. The first, second and third arguments above are similar to the (vertical and horizontal) arbitrariness, artificiality and fuzziness of introducing a geometrical rule that says that four-legged figures with outward pointing arrowheads decrease the length of line segments.

Also in the fourth argument, bodily appearance of a being is, just like the arrowheads, some external factor, and we have to universalize the rule that no external elements are important. In the Müller-Lyer illusion, this is the universalized rule that context is never important for determining a length. In daily life, we often (unconsciously) use the rule of context independence (you can simply ask an architect who makes a drawing of a house). It would be inconsistent to always use this rule, except in the case of the Müller-Lyer figure, because there is nothing really special about this figure. So we should apply context independence consistently. It would be strange that exactly in the Müller-Lyer figure context independence would not apply. That is why antiracists and antisexist should apply antidiscrimination consistently and hence become antispeciesists as well.

8.5 Five arguments in favor of sentience

In the previous part of this dissertation, I already addressed some arguments why sentience is morally relevant. In this section I summarize them again, and give a few more arguments. One of the reasons why there are different arguments for sentience is

that there are different moral virtues (empathy, impartiality, respect) and different normative systems. Different arguments for sentience stem from these different normative ethical systems and moral virtues. All the arguments have the same structure: starting with two assumptions (one factual and one value statement) one can derive that sentience is morally relevant.

1) Welfare ethics (consequentialism) and fairness ethics (contractualism):

Fact: Our own well-being matters to us.

Value: Impartiality is important. The thought experiment of the veil of ignorance (Rawls, 1971) is a nice tool to check impartiality. John Rawls only limited his theory to rational beings. But this thought experiment can be made more impartial (more consistent) when applied to all entities in the universe, as was proposed by Rowlands (1997, 1998). Imagine that you might be any object or entity in the universe, but you don't know who or what you might be. You could be a non-sentient thing without well-being, or a sentient being. How would you like to be treated? If you were non-sentient, this question would not matter to you, because nothing done to you will influence your well-being. You would not experience or prefer anything. So being sentient will imply a different treatment, because well-being matters to you⁵.

2) Virtue ethics and ethics of care:

Fact: We can feel empathy with all and only with sentient beings (beings who can feel and have a well-being).

Value: Developing the virtue of empathy (compassion) is important.

3) Rights ethics (deontology):

Fact: A sentient being is a being that has interests and can subjectively feel its interests. Feelings are nothing but affective conscious mental states that indicate that needs or interests are satisfied or not. For example pain indicates that bodily integrity is not satisfied, fear indicates that safety is not satisfied.

Value: Protection of interests by respecting rights is important. It is not farfetched to see a connection between rights, interests and feelings: feelings detect interests, interests are protected by rights. This is at least less farfetched than making a connection between e.g. rights and having certain genes, belonging to a certain biological group, or getting fertile offspring.

4) Other ethics:

⁵ Being human is not what would matter to you. To see this, ask yourself the question what you would prefer: you remain a human being but will be in a persistent coma without consciousness, or you turn into an animal but keep your mental capacities for well-being. I would prefer the latter, which means that I value mental capacities such as sentience more than biological categories such as species. Similarly, being male and being white is not what matters to me.

Fact: Mental capacities such as consciousness are something very complex and vulnerable in the universe.

Value: We should protect and respect entities that have vulnerable and complex mental capacities. Having a consciousness is at least something much more remarkable than having the genes of an arbitrary species. If a sentient being becomes a non-sentient being, he loses something valuable and does not gain something in return. On the other hand, if a white person becomes a black person, he loses one skin color but gains another; if a man becomes a woman, he loses one sexual organ but gains another; if a human becomes a non-human animal, he loses some physical properties and genes, but gains other.

5) The argument from marginal cases (Dombrowski, 1997):

Fact: Perceptual consciousness (sentience) is the only mental capacity that mentally disabled humans share with other humans.

Value: Our intuition says that mentally disabled persons are to be respected because of some inherent, mental capacity that they possess. The real reason why we help them is because they can suffer, they have interests, they can be harmed. Other reasons, such as indirect rights or a slippery slope argument made by Carruthers (1992), are in a sense disrespectful towards those individuals, because they deny their intrinsic value (see the heuristics argument in section 8.8). Neither is 'being alive' a sufficient criterion for giving mentally disabled persons rights, because human egg cells and embryos are also alive but they have a different moral status.

The above five arguments cohere with each other and indicate that sentience is a basis for moral concern and moral status. It is not farfetched to see a connection between rights, interests and sentience. This is at least less farfetched than making a connection between e.g. rights and the possibility of getting fertile offspring.

This set of five arguments is related to the translation invariance in the Müller-Lyer illusion. Just like length is an inherent property of a line segment, these arguments refer to a characteristic value of sentience. The first two arguments, which refer to impartiality and empathy, can be related to the idea that a ruler can be seen as a device to make our length judgments impartial (objective instead of subjective). As the ruler is a device to shift (translate) from one position to another, empathy and the veil of ignorance are (emotional and rational) devices that also help us to 'translate' ourselves into the positions of other sentient beings and measure how rich their emotional lives are (how important things are for them). Hence, impartiality in ethics is the analogue of translation invariance in geometry. Ethicists should develop compassion as a virtue, just as geometers should value the accuracy of rulers.

One might argue that the notion of sentience also has fuzzy boundaries, just like the notion of a species, as we have discussed above. When is a being sentient? What about invertebrates, plants,...?

This is first of all a matter of fact (science). As we've seen, scientists do not have and will never find indicators to determine at what point a being (a hybrid, a chimera, an ancestor or a genetically modified person) should be called human. But scientists already do have quite a lot of indicators to test whether a being is sentient (see next chapter). And they will likely discover new indicators when they gain more knowledge about how consciousness works. The species boundary has an inherent fuzziness; the sentience boundary is rather a matter of scientific uncertainty. A being cannot be both sentient and non-sentient at the same time. A being cannot both feel and not feel something at the same time. As with computers: either the 'sentience program' works (is switched on), or it doesn't.⁶

But a hybrid is half human and half non-human at the same time. A species always has an inherently arbitrary cut-off point. The boundary of a species is continuous, because a lot of properties that characterize a species are continuous properties. Hence, it is always arbitrary to select a point on this continuum. The cut-off point for sentience, on the other hand, is at the value of zero (i.e. the point where all positive and negative feelings become absent, where the 'feelings program' in the brains does not run). Such a zero point is (at least in theory) well-defined for sentience, but it is not well-defined for species. Look at all the ancestors of humans and ask the question: what exactly needs to be absent in order for an individual (an ancestor) to stop being a human being? This question cannot be answered in a non-arbitrary way.

Second, in our culture, non-human animals already have some moral status: look at the animal welfare laws. These laws refer to the welfare (sentience) of animals, so we are already able to use this criterion, even when there is still some scientific uncertainty about e.g. invertebrates.

Third, in human rights ethics as well there is scientific uncertainty about sentience: consider the discussion on abortion and stem cell research. There is scientific evidence that fertilized human egg cells are not (yet) sentient, so they have a lower moral status according to many people. Here also we are able to deal with this scientific uncertainty.

Fourth, even if there is an inherent gradation in the levels of sentience (from simple to complex emotions), it is not really a threat to the theory, because it makes sense to couple the gradation of sentience to a gradation of moral status (see the chapter on the basic right, section 6.4). All beings with a developed, complex, functioning central nervous system, all beings with a level of sentience equal or higher than those of (most) vertebrates, developed human fetuses or mentally disabled humans, have a very high moral status.

⁶ However, see the discussion about fractional number of minds (fractional consciousness) in section 4.6.8.

Fifth, if the sentience boundary is fuzzy, why add a second fuzzy boundary, the species? In ethics we should strive to avoid as much fuzzy notions as possible (otherwise we risk opportunism). We should delete the most arbitrary of the fuzzy boundaries: the species.

Referring to the Müller-Lyer illusion, we see that line segments have a continuous gradation (from short to long), but that is different from the gradation of the angles of the arrowheads. The former is an inherent property of line segments, the second is something external (contextual).

In summary, our current society has a patho-anthropocentric ethic. It takes two criteria into account: sentience and species. But it is better to drop the latter criterion, because the species boundary cannot determine or influence someone's moral status, just like arrowheads cannot influence the length of a line segment.

There is a very simple thought experiment that demonstrates that sentience, not species, is what matters. Imagine that tomorrow you will either remain a human being, but you will permanently lose consciousness, or you become a non-human sentient animal who is able to feel joy and other positive emotions. Which choice would you prefer? I would prefer the latter option.

8.6 Speciesism and cognitive impenetrability

As mentioned before, one characteristic of optical illusions is its cognitive impenetrability (Pylyshyn, 1999): even after measuring the lengths of the Müller-Lyer illusion, they still appear to be different. The question is whether speciesism has a similar kind of cognitive impenetrability: do our spontaneous, intuitive judgments regarding the moral status of humans versus animals still reflect some speciesism after we learn about the above arguments against speciesism? Although this question is difficult to answer, I am inclined to say yes, based on four reasons.

First, in discussions with meat eaters, a lot of people remain speciesist after learning about the above arguments. Those people give inconsistent counter-arguments (fallacies) to justify speciesism. This looks like a moral dumbfounding (Haidt, 2001; 2012), where people have strong intuitive moral judgments but fail to express a rational principle to explain their intuitive reactions.

Second, even some animal rights activists exhibit some speciesist language that reflects essentialistic thinking (see next section). For example in discussions those animal rights activists often refer to the notion of humans. This might indicate that those antispeciesist animal rights activists have difficulties in overcoming the moral

illusion. My personal experience confirms this: I am an animal rights activist, but I am aware that I still have some intuitive speciesist judgments about the moral status of humans and animals. It takes some cognitive effort to overcome those intuitions, just as it takes a cognitive effort for utilitarians to overcome the strong emotional intuition that we should not push a fat man from a bridge in order to stop a runaway trolley (Greene et al. 2004; Greene, 2008).

Third, it seems that a lot of antispeciesist animal rights activists have an emotionally different response towards eating human corpses versus non-human corpses. Eating dead human bodies (even if no human was killed and no human rights were violated) is accompanied with a strong feeling of moral disgust. A lot of animal rights activists do not have a similar strong feeling of disgust when it comes to eating dead animals (e.g. from road kill). It is unsure whether this difference in the feeling of disgust is related to a difference in moral status and whether it reflects a cognitive impenetrability of speciesism, but it might be an indicator.

Fourth, perhaps the best scientific evidence that speciesism is to some degree cognitively impenetrable, comes from studies on implicit associations (Greenwald et al., 1995; 1998; Devine, 2001). The Implicit Association Test (IAT) is an experiment to measure spontaneous implicit attitudes that people have towards e.g. races. The reaction speed is measured when experimental subjects have to associate pairs of concepts. Those concepts can refer to races (e.g. faces of black and white people) and values (e.g. positive and negative words like 'joy' and 'pain'). According to IAT studies, a lot of people have implicit racist attitudes, although those people have explicit antiracist attitudes: they can explicitly state that they are against racism and that they value black and white people equally, although they have shorter reaction times when they have to associate black people with negative values. Implicit prejudice and stereotyping might explain this difference between explicit and implicit attitudes (Devine, 2001).

Although an IAT-test about speciesism is not yet performed, I expect that those IAT-studies about racism and sexism can be extrapolated to speciesism. What these IAT-studies show, is a kind of cognitive impenetrability: even if a racist learns everything about racist prejudices and stereotyping, even if s/he recognizes how immoral racism is (that race is arbitrary, artificial and not morally relevant), his/her implicit negative attitudes towards other races do not simply disappear.

8.7 Psychological background theories: human prejudices and essentialism

One more thing needs explaining: what is the mechanism behind the moral illusion of speciesism? In the Müller-Lyer optical illusion, the coherent intuitions of context independence and translation invariance are brought into a ‘wide reflective equilibrium’ (Rawls, 1971) by introducing background theories about the underlying psychological mechanism. We know that the Müller-Lyer illusion is created by our brains, in order to adapt to 3D-vision. Our optical system has a bug; it’s stuck when looking at a 2D-image that reflects elements of 3D-perspective, such as the Müller-Lyer image. We have seen that our brains use a kind of heuristic (attribute substitution) to estimate lengths, using 3D-interpretations of e.g. staircases.

Also in the case of speciesism we have some well-established psychological knowledge about prejudice, stereotyping, the influence of language and words,... (see e.g. Plous, 2003, for some mechanisms behind prejudices towards animals). Let us summarize the psychological background theories that turn antispeciesism into a wide reflective equilibrium.

In-group-out-group bias

Psychologists studied the mechanisms behind optical illusions such as the Müller-Lyer illusion (Purves & Lotto, 2002). Also a lot of research has been done on the psychology of discrimination, focusing on e.g. stereotyping and prejudice (Whitley & Kite, 2010). In-group-out-group discrimination is based on a cognitive bias: in-group-out-group bias or in-group favoritism (Tajfel, 1981; Whitley & Kite, 2010) refers to a pattern of favoring one’s in-group members over out-group members. This bias contains elements such as out-group homogeneity (Quattrone & Jones, 1980; Rubin & Badaea, 2012), a pattern of underestimating the differences between out-group members. Also in the case of speciesism we have some well-established psychological knowledge about prejudice, stereotyping and the influence of language and words (see e.g. Plous, 2003).

Essentialism

Our brains appear to be trained in essentialist thinking to categorize groups. The first three arguments I presented against the species boundary indicate that there really is no essence related to a species (see also Hull, 1986). Essentialism means that there are characteristics that all elements of a specific set (e.g. a species) possess and elements of other sets don’t possess. All elements of that set can be accurately described and defined by those characteristics. That specific set therefore has a unique definition.

Essentialism in biology is rooted (or reflected) in ancient philosophical thinking (e.g. Platonism), as well as major religions. In those religions it is believed that there is something special to all and only to humans: all humans, and only humans, have an eternal soul, or are created in the image of God. But since Darwin, the scientific consensus says that there is nothing special about a species. It is just an arbitrary abstract classification with its limitations and difficulties (Rachels, 1990). Similarly, a racist thinks of races or ethnic groups as being essentialized natural groups, even though it is now well known that there really is no essence related to an ethnic group or race.

Even more: several studies give explanations for this phenomena that people rapidly (but incorrectly) tend to categorize entities in terms of essentialized groups (Gil-White, 2001). Our intuitions are not always in line with science. According to Gelman (2003) and many other psychologists, children and adults intuitively describe biological entities in essentialist terms. People (from different cultures and backgrounds) automatically think that biological categories have invisible essences (Bloom, 2010). As we have seen in the section on discrimination in part two of this dissertation, the psychological mechanisms of hierarchic dualisms also tend to work with essentialistic concepts: e.g. the big gap between the upper and lower side, the homogenization of the lower side,...

Looking at the literature, it is remarkable how many people defending speciesism are essentialists, by referring to personhood or humanity (the human species) as a 'kind', having a 'substantial nature'. (See e.g. Chappell, 2011; Cohen, 2001; Finnis, 1995, p.48; Lee & George, 2008; Scanlon, 1998, p.186; Scruton, 2000. See McMahan, 2005 and Tanner, 2006, for an extensive critique of the 'argument of kinds'.)

This subtle mechanism is also reflected in our language. It is amazing how often one encounters human-centric notions in our culture without even noticing. Look again at the title of this section: what is the word 'human' doing there? Why not 'primate'? Or look at definitions of discrimination: how often are these definitions already from the start restricted to humans (or persons, where a person automatically means a human)? If people might restrict the definition of discrimination to our species, then a racist is allowed to restrict this definition to whites. It is better to start from a really impartial definition, as we have done in the chapter about discrimination (section 7.3).

Look at discussions between speciesists and vegans. How often do speciesist people respond with: "But humans..."? 'The human' does not exist. If it exists, then 'the primate' or 'the mammal' would also exist. Why does no-one mention them? Antispeciesists who grew up in a speciesist society, really have to 'deprogram' themselves. Often animal rights activists still use some essentialistic language. Once you are completely deprogrammed, you start to see how strange this constant referring to 'humans' really is. It sounds like someone is constantly referring to 'dry-nosed primates'. If you hear someone saying 'humans and animals', it sounds as crazy as

‘primates and animals’. The reader is invited to read any book on animal ethics, and to replace everywhere ‘humans’ into e.g. ‘primates’ or ‘placental mammals’.

So if in a discussion about animal rights people respond by saying something like: “But most humans have rationality and the capacity of moral thinking”, the very same statement would be true for the family of great apes or perhaps also primates. Most great apes also have rationality (just count them: more than 6 billion great apes have high levels of rationality).⁷ And if species has an essence (of say rationality), why should the class of mammals not have an essence either? It is equally possible to look at mammals as a kind. If mentally disabled persons get special rights because they belong to the kind of humans with a rational nature (most humans have rationality), then it is also fair to say that humans do not get special rights, because they belong to the kind of mammals, lacking a rational nature (most mammals are not rational beings). Our language is strongly biased towards one group (the species of humans), by presenting this group as having a substantial nature or a kind. Meanwhile, it neglects the other possible groups, natures or kinds.

Does language simply reflect our essentialist thinking, or is our essentialist thinking amplified by our language? In any case, the fact that a speciesist tends to think of species as essentialized groups does not imply that there is an essence to a species.

Let’s briefly refer back to the Müller-Lyer illusion. We can say that straight lines have an important essence, as they are primitive geometrical objects or can be defined in a unique, simple way (e.g. zero curvature, shortest distance between points). But it would be strange to speak of an important essence related to all line segments having outward directing arrowheads. These figures do not form an important category of geometrical objects, because they are vaguely and arbitrarily defined. It requires a lot of information to correctly define such geometrical objects, just as it requires a lot of information to correctly define a species such as *Homo sapiens*. And many elements in those definitions will appear highly arbitrary. We built a geometry on lines; we do not built a geometry on special figures with arrowheads. Similarly, we should built an ethic on morally relevant criteria (e.g. well-being), not on highly complex categories such as species.

There is more to say: essentialism is a very clear example of a heuristic that uses attribute substitution (see Sunstein, 2005). The next section gives an extensive discussion on speciesism as a prime example of a moral heuristic.⁸

⁷ Interestingly, after hearing this statement, people often have an automatic response that most of the great apes do not have rationality: only humans have rationality, whereas orang-utans, gorillas, chimpanzees and bonobos do not. That is four against one. This is other evidence that people tend to think in terms of species (or genera, because there are different species of gorillas) instead of individuals.

⁸ The section is based on Bruers (2013a).

8.8 Speciesism as a moral heuristic

Ever since Ryder (1971) introduced the term speciesism – a prejudicial discrimination on the basis of species membership – more than 40 years ago, it has attracted a great deal of controversy⁹. This chapter combines the philosophical reflections on speciesism with a recent development in moral psychology, namely moral heuristics (Sunstein, 2005).

Do the test: ask any person what justifies our current use of animals for experiments, food, clothing or entertainment. Chances are high that you will hear an answer that sooner or later refers to a distinction between humans and non-human animals. Next, you can ask them what it is about humans that other animals lack and that justifies a different treatment of humans and animals. Most people will answer this question, so it is a common belief that this is a meaningful question. Now, again the chances are very high that the answer will refer to a mental capacity that most humans have and animals lack: self-consciousness, creativity, rational reflection, the ability to speak, understand ethics, sign social contracts, have a sense of justice, and many others. The list of authors and philosophers who have defended speciesism by referring to such mental capacities is long (see e.g. Carruthers, 1992; Cohen, 1997; Scruton, 1998).

The antispeciesist now comes up with the ‘argument from marginal cases’ (see e.g. Dombrowski, 1997; Wilson, 2001), which might be better (more neutrally) termed ‘argument from atypical humans’. Atypical humans refer to a minority group of *Homo sapiens* who lack mental capacities such as rationality. The argument says that such atypical humans exist, and giving those atypical humans a moral status comparable to typical humans would be inconsistent if the mental capacity is a necessary condition for moral status.

Confronted with this argument from atypical humans, some people defending speciesism attempt to extend or refine their criteria in the hope of including all atypical humans (and still excluding all non-human animals). They refer to the potentiality of developing a certain mental capacity in the future, the possibility that they themselves might later become mentally handicapped, or the presence of interpersonal relationships between those atypical humans and typical humans.

However, it is striking that those attempts are too often doomed to failure (some authors who have defended the argument from atypical humans against such attempts include: Dombrowski (1997), Huther (2005), McMahan (2005) and Tanner (2006, 2009)). The antispeciesist can persist by referring to more extraordinary atypical humans who

⁹ For some recent discussions in the literature, see Bernstein (2004), Chappell (2011), Grau (2010), Horta (2010b), Lee and George (2008), Liao (2010), McMahan (2005), Nobis (2004) and Tanner (2009).

fall outside the scope of those extended and refined criteria. The abovementioned criteria are invalid when applied to for example an incurable, seriously mentally handicapped young orphan. Such human beings exist (in fact, I happen to be a foster parent of such a Vietnamese boy). I am not aware of any proposed set of mental capacities plus refinements that allows the inclusion of such humans in the moral realm, and at the same time excludes all non-human animals.

In the many conversations that I have had, my opponents who defended speciesism often gave one final response: a simple affirmation that those extraordinary atypical humans are still humans and therefore should be protected. Strikingly, people giving such a response are often not aware of the circularity in this reasoning. And what is more: using the argument from atypical humans in such conversations often triggered reactions varying from indignation to overt outrage.

8.8.1 The heuristics hypothesis

For animal rights advocates, the above sounds very familiar. The hypothesis that I want to put forward is that this common speciesist thinking is based on a heuristic. Heuristics are intuitive, efficient rules of thumb applied when facing complex problems (Kahneman & Shane, 2002). As will become important in our discussion of speciesism, these heuristics work by a process called ‘attribute substitution’: our brains (unconsciously) substitute a computationally complex target attribute for a heuristic attribute that is easier to calculate or detect. In recent literature, as a spin-off of the work of Kahneman and Tversky (1982), the study of moral heuristics has gained some influence (Sunstein, 2005; Sinnott-Armstrong, Young & Cushman, 2010). In general, a heuristic works pretty well in most cases, but as Sunstein argued, in certain, atypical situations, moral heuristics might ‘misfire’ and create erroneous intuitive judgments. I am going to argue that this misfiring of the heuristic is exactly the case in situations with atypical humans. In fact, speciesism is a very clear example of the mechanism of attribute substitution. If the speciesism heuristic hypothesis is true, it can explain why a lot of people are ‘blind’ to the argument from atypical humans, why a lot of people do not seem to be aware that they deny the rights of mentally disabled humans when pointing at some complex mental capacities, and why the speciesism intuition is so obstinate.

A lot of people have the conviction that moral status depends on a complex mental capacity, such as rationality. This mental property is the so-called target attribute of a being. But the problem of this target attribute is that it is difficult to detect. If we encounter a being, how can we quickly decide whether or not she has the relevant mental property? Our brains have found a solution: they unconsciously substitute the target attribute for a heuristic attribute that is easier to detect. This heuristic attribute

is based on something our brains are good at: pattern recognition (Margolis, 1987). For example, looking at figures, we can very quickly interpret a figure as the letter A, without being able to explain what exactly characterizes a letter A. Computers are not (yet) able to detect a letter so quickly. Similarly, looking at objects, we can very quickly determine whether it is a human, even if no-one is able to clarify what set of elements, conditions and characteristics defines a human being. We look at an individual and immediately see the pattern (face, behavior, etc.) that corresponds to a human, because our brains are trained that way. Now, looking at the set of objects that have the target attribute of rationality on the one hand, and the set of objects that have the heuristic attribute of a human being on the other hand, we see a strong overlap between these two sets, with a low percentage of exceptions. The exceptions are the atypical humans. Most beings that have the 'human pattern' also have the target attribute. So our brains use the species criterion (our human recognition capacity) as a heuristic. When an object looks like a human, when it has the characteristic pattern of a human, intuition says that the object has the target attribute mental capacity as well. This 'speciesism' or 'looks-like-a-human' heuristic works pretty well in most cases, but not in the atypical cases. If speciesism is a heuristic, it explains why antispeciesist people so often refer to the argument from atypical cases.

The speciesism heuristic becomes particularly clear in the recent work of, among others, Chappell (2011), who refers to a 'folk' notion of human species (which – as I interpret it – is based on our pattern recognition skills) to determine who counts as a person: "In normal cases, we have already identified a creature as a person before we start looking for it to manifest the personal properties, indeed this pre-identification is part of what makes it possible for us to see and interpret the creature as a person in the first place. And that pre-identification typically runs on biological lines." (Chappell, 2011, p.1). The pre-identification is nothing but the attribute substitution: our brains immediately and unconsciously substitute the target attribute (a property of personhood – or what Chappell and others might have in mind: a complex pattern of mental properties that constitute personhood) for a heuristic attribute (that "runs on biological lines").

Even though we know that heuristics can sometimes result in erroneous intuitions or judgments, it does not imply that we are better off without heuristics. Sunstein (2005) and rule utilitarians (see discussion in Shaw, 1999, pp.145-170) argued that without those heuristics or rules of thumb we might make more mistakes. The question I address in this section is whether using the speciesism heuristic is permissible, useful or dangerous. Do we have to keep it, improve it or throw it away because it makes some errors in atypical cases?

In the following sections I discuss the strongest pros and cons of keeping the speciesism heuristic in atypical cases. Afterwards, I argue that – even if it is not irrational or inconsistent to stick to the speciesist heuristic – it is better (more

respectful towards atypical humans) to take another heuristic which uses sentience instead of rationality as its target attribute.

8.8.2 Time and knowledge constraints

One advantage of heuristics is that these are rules of thumb that can help us make quick decisions in situations with time and knowledge constraints. Compare heuristics with traffic laws. The target attribute in traffic would be a rule such as: “Always drive as to maximize well-being” or “maximize efficiency and minimize accidents.” This target attribute rule is too difficult to follow, so it is substituted for simpler heuristic rules, such as: “Always stop in front of a red traffic light.” But in atypical cases, when there really is no other traffic around; there is no harm in ignoring a red light. Most people would say that introducing a new traffic law: “Stop at red lights except when crossroads are safe,” would make matters worse, because we cannot be sure enough whether crossroads are safe. Perhaps we are not smart or alert enough to judge the safety. Perhaps we are tempted to judge safety to our own advantage. Perhaps we are biased and we ignore red lights even when the situation is not safe.

So, it is often conceded that strongly holding on to heuristics is a good strategy. Does the same apply for the speciesism heuristic? The difference between traffic situations and situations related to treatment of atypical humans is that in the latter we do have time and (scientific) knowledge to influence our decisions. True, in emergency situations, the analogy with traffic might be valid. If you see some creature drowning, and you have to be quick to decide to rescue that being, it would be effortful, time-consuming and unreliable to look first for the mental capacities of that drowning being. If you see it is a human, you will show a direct response and jump in the water to save this human. And in most cases, your judgment will be correct: in most cases, the drowning human will be a rational self-conscious being who deserves to be saved. Weighing the probability that it is a mentally handicapped human against the cost for you to rescue the human, would still make you conclude that it is better to stick to the speciesism (‘looks-like-a-human’) heuristic.

But when we have to decide how to treat mentally disabled persons, whether we should clothe them, feed them or use them in experiments, we do have time and access to information about their mental states. In these cases, other heuristics than the ‘looks like a human’ heuristic might be more accurate, in the sense that these new heuristics also cover all rational beings, but include fewer non-rational beings. For example, we could look at results of communication or IQ-tests, adaptive behavior or neurological functioning. And scientists might come up with more accurate and faster techniques to see what the mental capacities of a mentally disabled human are. People with average intelligence might be vulnerable to bias and erroneous judgments about the mental

capacities of beings, but scientists and judges might be able to make sufficiently wise judgments.

So where do we stand? Do we feel comfortable with the idea that having fundamental rights would depend on our subjective state of knowledge? It is absurd to claim that mentally disabled humans have rights merely because we are at this moment ‘too stupid’ to work with more accurate heuristics. We, and at least scientists, philosophers, and judges, are intelligent enough to determine which human being is certainly not a rational being, and I am not aware of historical or psychological evidence that suggests that using the target attribute directly or using more accurate heuristic attributes instead of the speciesist heuristic attribute results in real violations of the rights of rational beings. So, in most cases we do have time and we have already developed efficient ways to detect mental capacities. But one might object: how reliable are those scientific tests? And a more fundamental question is how reliable should these tests and refined heuristics be? Answering that question eventually becomes a matter of taste, of gut-feelings. Although my intuition says that time and knowledge constraints are not sufficient reasons to stick to the speciesism heuristic, we should accept that this discussion remains unresolved and that it is not yet irrational or inconsistent to stick to the speciesism heuristic due to the above concerns about our limitations of knowledge.

8.8.3 Fear of a slippery slope

Following Carruthers’s slippery slope argument (Carruthers, 1992), one could argue that it might be better to retain the speciesism heuristic. Several philosophers have criticized the slippery slope objection of Carruthers (Dombrowski, 1997; Tanner, 2009), but we can look at this argument from the heuristics perspective.

People might worry that not retaining the speciesism heuristic, i.e. using the target attribute of rationality directly instead of the heuristic attribute, might result in more serious errors overall. More rights of real rights holders (i.e. truly rational people) might be violated, because the target attribute is difficult to detect or there is no sharp distinction between having and not having the target attribute.

A first reason that people might give to justify this view is that rationality and other mental capacities are a matter of degree. So we have a ‘sorites’ problem of where to draw the line. When removing grains from a heap, when does the heap become a non-heap? When removing mental features, when does a person lose its rationality? If we cannot answer this question, we risk making erroneous judgments about the rationality of some persons.

Second, one might point to the fact that our cognitive biases can unconsciously skew our judgments. In situations involving possible atypical humans, we might be vulnerable to bias and erroneously judge the situation to our own benefit (we might too

easily start to think that a specific human is non-rational, and treat him or her as a non-rational being to our own benefit).

These two observations combined will put us on a dangerous slippery slope, where we will move towards real violations of the rights of rational people. But also two objections to this slippery-slope argument can be raised. The first is that, if there is indeed an unavoidable spectrum of mental capacities, we might be able to couple this to a spectrum of rights. Some people with higher mental capacities could be given more or stronger rights claims.

However, this first objection might not run so smoothly. When it comes to fundamental rights, some might prefer to stick to the binary view: either one has an absolute claim to this right, or one does not have the right at all. Such a binary view cannot be coupled in a non-arbitrary way to the supposed spectrum of mental capacities.

A second counter-argument to the slippery-slope argument is that we are already able to deal with such slippery slopes. Consider situations where we have to decide whether mentally handicapped humans have a right to vote or a right to marriage. Different countries have different ways of dealing with the right to vote for mentally handicapped humans. In some countries, a judge or medical practitioner will decide whether a mentally handicapped person has this right. In others, the person needs to undergo a psychological test or needs to be under a protective measure such as a guardianship (for the situation in European countries, consult FRA (2010)). Whatever solution a country prefers, there seems to be a general lack of worry that this exclusion of mentally handicapped persons from the right to vote would put us on a slippery slope towards broad violations of the right of rational humans to vote.

However, this second counter-argument is based on a presupposed analogy between a right to vote and a more fundamental right such as the right not to be harmed or the right not to be used as merely means to some else's ends. One might object that we should be more concerned about slipping down a slippery slope when the slope involves a fundamental right. A second possible objection to this second counter-argument is that the demarcation line between people who can have the right to vote and those who do not might be easier to draw than the demarcation line between rational and non-rational people. It might be easier to check whether someone is able to vote (i.e. by doing a communication test), than to check whether someone is able to reason or is self-conscious. The analogy between spectra of fundamental rights and spectra of political rights might be too weak.

The above discussion indicates that, as with the argument of time and knowledge constraints, things are not yet completely resolved. It might come down to a kind of uncertainty aversion: if people have a strong fear for slipping down the slope of fundamental rights when we are confronted with atypical humans, they have a strong uncertainty aversion. They are worried about the question: "What might happen to my

rights and the rights of my loved ones, if we stop giving fundamental rights to mentally handicapped humans?”

The major problem that I have with this slippery-slope defense of the speciesism heuristic is that it seems disrespectful to claim that the moral status of mentally handicapped people merely depends on our uncertainty aversion or our supposed inability to put barriers on a slippery slope, instead of it depending for example on the real interests and feelings of those mentally handicapped people. Are we so sure that we will slip down the slope when we look for more accurate heuristics than the speciesism heuristic, say a heuristic based on some psychological tests? This question has no easy answer, but at least to me, refinements of the heuristic rule (to make it better fit with the target attribute such as rationality) do not seem to be impossible, nor do they seem to be so dangerous for the rational people. They are dangerous for a-rational, atypical humans.

Also, I am doubtful that those people who defend the speciesism heuristic due to an uncertainty aversion, have a consistently strong uncertainty aversion in other situations in their lives. It seems strange to me that merely avoiding a slippery slope is the real motivation for people to take such care of mentally disabled orphans. If that were the real motivation, we could expect that one would have a very high level of uncertainty aversion and fear of slippery slopes. But such a high level of uncertainty aversion seems incompatible with the way we deny some atypical humans a right to vote (it is unlikely that judges are really unable to make wise decisions about who is able to vote), and with the treatment of animals in factory farms. We do not seem to worry at all about slippery slopes or the potential negative influence on our rights when we treat thousands of sentient animals the way we do in factory farms. But, admittedly, it is difficult to test such apparent inconsistencies in people's uncertainty averse attitudes towards slippery slopes.

8.8.4 The emotional cost of excluding atypical humans

Moral heuristics are often strongly internalized rules, which means that rule violations are often accompanied with strong emotions of indignation, guilt or moral disgust. The abovementioned reactions of people (sometimes overt outrage when they are confronted with the argument from marginal cases) indicate that the speciesism heuristic and the rights of atypical humans are also strongly emotionally charged. People have empathic concerns for the mentally disabled. Even if the emotions that people feel towards atypical humans would be irrational if all that mattered was a property (mental capacity) that those atypical humans lack, we observe that violating the heuristic will result in an emotional cost, and this cost is not to be underestimated. Sometimes it might be rational to stick to irrational feelings.

Compare it with fear of heights. Imagine that most people had a strong fear of heights. Do they react irrationally? Not necessarily: these people might claim that, if they conquered their fear in situations where they could not fall (e.g. when they are safely attached), they might react less fearfully in more dangerous situations where fear is required or advantageous. They might know about themselves that they will be tempted to make erroneous judgments in dangerous situations. And they also know that it takes some effort (e.g. some costly therapy or focussed meditation) to overcome their fear of heights. So these people have a heuristic: always avoid tall buildings. The costs of overcoming their fear might be greater than the cost of avoiding tall buildings, so even when they could not fall from the tall building, they do not necessarily react irrationally by keeping the heuristic. These people have weighed all the costs and benefits, including the emotional ones.

The same could be said about the feeling of indignation that one experiences when looking at human rights violations, even if the human is mentally handicapped and is lacking the relevant mental capacities. The emotional cost of reacting in a more detached or neutral way towards those atypical humans might be greater than the benefits that one could obtain from violating their rights.

However, this weighing of the emotional cost against the potential benefits is often very difficult. One might object that the benefits for real rights holders of violating the rights of atypical humans should not be underestimated either. There might be health advantages in using atypical humans in e.g. medical experiments. Atypical humans (such as mentally disabled orphans) are often better research models for rational humans, compared with non-human animals, because the atypical humans are genetically and physiologically closer to the rational persons. So their use in medical experiments might give better results than if non-human animals are used as models for rational persons. The atypical humans could also be used for organ transplantations and blood transfusions, to help rational persons in need. If we do not sacrifice non-rational, atypical humans, then rational humans might die. Is that not more serious than the abovementioned emotional cost? It is not easy to decide this issue.

Consider again the traffic laws. Most people feel repugnance when they drive through a red light, because they have internalized an important rule. But are we not allowed to drive through red lights in emergency situations (e.g. when there is a child in the middle of the crossroad, in the distance there is a car coming, and we could only bring the child to safety by ignoring the red light)? How stubbornly do we have to stick to the heuristic traffic rules ("always stop at red lights") in such situations? As with some of the previous questions that I raised, we have to admit that these questions do not have easy answers.

For me, the 'emotional cost' defense of the speciesism heuristic seems to be too weak at this moment, but I again have to admit that my judgments might be biased, that the emotional cost is not to be underestimated either or that using atypical humans in

experiments (instead of using only non-human animals) would not be a sufficient improvement for the health of rational beings.

In summary, the above three defenses for the speciesism heuristic (time and knowledge constraints, slippery slope, and emotional cost), remain largely undecided and are neither clearly irrational nor inconsistent. In the next and final section, I will explore my main objection to the speciesism heuristic, even if this speciesism heuristic was not applied in an irrational or inconsistent manner.

8.8.5 The importance of sentience

The speciesism heuristic was based on the assumption that the target attribute is some higher mental capacity such as rationality. In my view, this seriously underestimates the importance of another mental capacity that even most animals have: sentience. The importance of sentience can be seen by asking the questions: what is the real reason why people help mentally handicapped humans in institutes? What really drives those health care workers to take care of atypical humans? I do not believe that they are willing to accept that the only reason why they take such great care of the mentally disabled is that they fear a slippery slope or the emotional costs of overcoming a heuristic.

The idea that some humans have rights merely because our heuristic misfires seems incompatible with the moral intuitions of many people. For example, it is more plausible that the persons who take care of handicapped people in fact respond to the needs of these people. They are not concerned with a higher mental capacity as the morally relevant target attribute. These health care workers have empathy, and they are happy when they see that the mentally disabled humans feel pleasure or joy in something. They want to avoid their suffering. Therefore, according to these carers, sentience is one of the most important target attributes, and sentience is likely to be the most important motivator for them to help these mentally handicapped people. If the empathy and moral intuitions of a health care worker towards atypical humans are clear expressions of an undercurrent in the common morality of our culture, which I believe they are, we can say that sentience is important in our common morality. But its importance is underestimated due to the dominance of the speciesism heuristic.

The problem with the speciesism heuristic is that it claims that rights of atypical humans are only indirect results of misfiring heuristics and that these humans in fact do not deserve rights because they are not rational beings. But it is highly disrespectful towards atypical humans to say that they only have an indirect moral status, that they in fact do not deserve rights but that we intuitively give them rights merely because our speciesism heuristic misfires.

Consider the mainstay of the speciesism heuristic: the ability to see a difference between humans and non-human animals. Without such a clear observable distinction, the heuristic attribute would not be that useful. But as biology now shows, this human-animal difference is not absolute or essential. In the past, there were human ancestors with more and more non-human (non-rational) properties as we look further into the past. There is in fact a whole continuum of ancestors, moving down the evolutionary branch, till we meet a common ancestor of, say, humans and pigs. Also, it might not be genetically impossible for human-chimpanzee hybrids to be born. Such hybrids are infertile offspring of a human and a chimpanzee parent, which means that each cell contains the DNA of both humans and chimpanzees. Or what about human-animal chimeras; beings who consist partly of human body cells, partly of non-human body cells? Or what about genetically modified humanlike beings? It is hard to believe that we really feel comfortable with the thought that, if the ancestors, hybrids, chimeras or genetically modified humanlike beings were alive among us, our speciesism heuristic loses its strength and we would drastically alter our ethics and our treatment of atypical humans. Although these examples are hypothetical, they should give us some discomfort.

It is awkward to claim that mentally disabled humans are just lucky that we have pattern recognition skills and that for us in the current situation it is easier to see distinctions between humans and non-humans than to see distinctions between rational and non-rational beings. Disabled humans have basic rights, but not because they are just lucky that the borderline human/non-humans do not exist yet or do not exist anymore. And neither are they just lucky that the borderline rational/non-rational beings do exist. If there were no borderline cases of rational/non-rational humans, if we were clearly able to make a demarcation between rational and non-rational humans, the slippery slope argument would completely fail, because the slope would contain a really big gap, and this gap is a good place to stop any further slipping down the slope.¹⁰

¹⁰ Mentally disabled humans would only have an indirect or dependent moral status if their moral status depends on the existence of human beings with intermediate levels of rationality and the non-existence of human-animal hybrid beings. Hence, a dependent moral status violates the intuition of independence. This can be compared with the discussion on independence in the prioritarian welfare ethic (see section 4.6.7.1). Whereas I could tolerate a violation of independence in the prioritarian welfare ethic (where the level of priority for someone's well-being can depend on the (non)existence of other beings), I do not tolerate a violation of independence when it comes to such a fundamental aspect as someone's moral status. As we have seen, the violation of independence in the welfare ethic is justified on the grounds of coherence between two strong intuitions: the preferences of an impartial observer behind the veil of ignorance and the avoidance of the repugnant conclusion together make a strong case to allow violations of independence. The violation of independence in the case of granting moral status cannot be justified because a similar strong coherence is lacking.

All the above questions and reflections should take us to the conclusion that sentience is more important than rationality. Sentience is a better target attribute from a moral point of view, for two main reasons. The first reason is based on impartiality. Rowlands (1997, 1998) and Van den Berg (2011) derived the sentience criterion through a contractarian ‘veil of ignorance’ thought experiment. This thought experiment tests our impartiality by forcing us to take the positions of others, as if behind a veil of ignorance we do not know whose life we are going to live. The veil needs to be as thick as possible in order to respect maximum impartiality. This means that we have to include the positions of all non-humans, non-rational beings and non-sentient beings. If I put myself in the position of a non-sentient being, things would not matter to me, because I would not have any subjective experiences or consciousness. If I were a non-rational but sentient being, things would matter to me, because I would still have a sense of well-being. This reference to preferences and what would matter to the subject is also the basis of a utilitarian-consequentialist vindication of sentience (Singer, 1975). In summary, when we value our own well-being (what matters to us) and we value impartiality as in consequentialist or contractarian ethics, then the well-being of all sentient beings should be valued.

This impartiality argument for the sentience criterion is also coherent with a second argument in favor of sentience: the virtue of compassion (Slote, 2001). In almost all major religious and philosophical traditions compassion is considered as one of the greatest of virtues. It is based on a feeling of empathy for the suffering of others and a desire to act on that emotion. Compassion is directly related to sentience because we are able to feel empathy with (non-rational) sentient beings, but not with non-sentient beings who cannot suffer. Compassion also plays a key role in an ethics of care (Gilligan, 1982). Compassion is what drives health care workers to help mentally disabled people.

Impartiality (supported by some interpretations of consequentialist and contractarian ethics) and compassion (supported by some interpretations of virtue ethics and ethics of care) both point to the importance of sentience. Such a coherent justification for sentience is lacking for the higher mental capacities such as rationality. From an impartial perspective, having a sense of well-being is not restricted to rational beings alone. And from a virtues perspective, there is no moral virtue that restricts attention to rational beings alone¹¹. There are two arguments to justify rationality, but these arguments can also justify sentience.

The first argument to justify the criterion of rationality is by pointing out that a coupling between rights and rationality is not far-fetched: rationality can be defined as

¹¹ Although one can argue that virtues like honesty and fair-mindedness indirectly refer to a notion of rationality. However, the virtue of compassion directly and strongly refers to suffering and sentience.

the ability to understand and respect rights and interests. But neither is the coupling between rights and sentience far-fetched: rights protect interests and feelings detect interests. For example, pain indicates a violation of bodily integrity and fear detects the interest of safety. Sentient beings value their own interests due to their positive and negative feelings. Therefore, it is not far-fetched to couple rights to sentience. From a rights perspective, both rationality and sentience are equally valid.¹²

A second way to justify the criterion of rationality is by referring to intuitions that some people have towards rational beings. Some people simply intuit that rational beings have a higher moral status than other beings. However, the sentience criterion is also coherent with our intuitions about helping mentally disabled humans or pet animals such as dogs, and intuitions about preferring animal welfare laws. An exclusive focus on rationality in ethics cannot explain those attitudes towards non-rational beings.¹³

As with other mental capacities, sentience is difficult to detect. Of course, when confronted with an individual being, we can always try to do some tests to see whether it is sentient. If we take sentience as the most important target attribute, and if sentience is difficult to detect, we can look at a suitable heuristic attribute. Heuristics might be useful in many cases, so we should not throw away all such rules of thumb. Looking at our current scientific knowledge about sentience (see next chapter), we can take the biological group of vertebrates as the corresponding heuristic attribute: vertebrates are also easy to recognize, and science indicates that there is a strong overlap between the group of vertebrates and the group of sentient beings (see e.g. Griffin (2001) and EFSA (2009) for sentience in fish).

I suggest that – in contrast to our rather ‘fixed’ attitude towards the speciesist heuristic – we can and should have a more ‘flexible’ attitude towards the vertebrate heuristic. When possible (when scientists have accurate ways of determining sentience), it might be best to dispense with all heuristics, including the vertebrate heuristic. For example, we might also have to include some large crustaceans and molluscs such as squid, because they also might be sentient. And presumably some atypical vertebrate animals are non-sentient.

It would not be disrespectful towards those non-sentient animals if we do not stick to the vertebrate heuristic, i.e. if we were to give those specific non-sentient individuals a lower moral status than sentient vertebrates. The reason why the speciesist heuristic was disrespectful towards non-rational humans is exactly because those humans are

¹² The coupling between rights and species is really far-fetched, because we cannot see a connection between rights and genes or between rights and the ability to beget fertile offspring.

¹³ In common morality, the attitudes towards different non-rational beings are not consistent. Consistency can be improved by uplifting the moral status of non-human vertebrate animals, due to their sentience.

sentient: they have a sense of well-being and feelings that express interests. The disrespect does not lie in the use of a heuristic, but in the choice of the target attribute. Choosing a target attribute that excludes sentient but non-rational humans is disrespectful, even when these atypical humans are saved by the ‘misfiring’ of a speciesist heuristic.

To conclude, people give too much credence to the arguments in favor of the speciesism heuristic, or to arguments in favor of speciesism in general. The reason why they give too much credence to those arguments is perhaps because those people use animals on a huge scale. For example, they decide three times a day to eat animal products. They have friends and family who use animals in similar ways. They know that their parents and grandparents used animals in similar ways. They see TV commercials that promote meat, see animal circuses in their hometowns, and see no-one (or just a few ‘extremists’) complaining, etc. For those who consume animals on a daily basis, a lot is at stake (especially for their self-image), so we can expect that this creates a real bias towards justifications of speciesism and the use of animals. Those people are less willing to accept the moral importance of sentience, and the extension of rights to all sentient beings.

8.9 Summary

Looking at the above, we now have a fairly strong coherent picture that implies that discrimination such as speciesism is a moral illusion. Its vertical and horizontal arbitrariness, its artificiality, its violation of impartiality, its cognitive impenetrability, its relation to cognitive biases such as essentialism and heuristics, and the fact that ideologies such as white-dominant racism, male-dominant sexism or human-dominant speciesism are strongly culturally determined, all corroborate the conclusion that discrimination is an illusion.

I presented a set of five arguments why the species boundary is irrelevant, and another set of five arguments why sentience is relevant. The first set of arguments (against the species boundary and essentialist thinking) is analogous to the principle of context-independence in the optical illusion. The second set corresponds with the principle of translation invariance in geometry.

All these arguments cohere with each other: we have a situation of narrow reflective equilibrium (Daniels, 1979) where strong intuitions and principles mutually support each other, and according to this narrow reflective equilibrium, speciesism is an illusion.

But there is more: speciesism is also in conflict with a wide reflective equilibrium (see Daniels, 1979 for the notion of wide reflective equilibrium). This wide reflective equilibrium not only contains moral intuitions and ethical principles, but also contains (scientific) background knowledge about e.g. psychology and cultural anthropology. We know that not everyone is susceptible to speciesism (it depends on culture and education), and more importantly: we do have insights in the psychological mechanisms behind speciesism: heuristics (attribute substitution) and essentialist thinking (using language with prejudices, stereotyping,...). Compare this with the Müller-Lyer illusion: we know that not everyone is susceptible to this illusion (Segall et al. 1963), and we do have insights in the optical mechanisms behind this illusion (a heuristic of automatic perspective corrections from 3D to 2D).

In geometry we have a very coherent picture that is much stronger than this one optical intuition about the differences in lengths of the Müller-Lyer figure. I therefore believe that the whole antispeciesist picture is coherent to such a high degree that it is much stronger than that one moral intuition about the moral status gap between humans and non-human animals. What else would the speciesist need in order to be convinced? That the speciesist intuition is cognitively impenetrable or that essentialist thinking happens automatically, are not sufficient reasons to say that the species boundary is morally relevant.

At one point the analogy between the Müller-Lyer illusion and the speciesism illusion goes wrong: the huge difference in our treatment of pigs versus mentally disabled humans can only mean that there is a huge effect of belonging to the human species. On the other hand, the Müller-Lyer illusion is rather subtle; it does not create huge differences in length judgments.

Nevertheless, I expect that this new way of looking towards speciesism can shed a light on why the speciesist intuition is so pervasive and difficult to change. As optical illusions, the speciesism illusion is cognitively impenetrable. The analogy with optical illusions might help us to argue for a more consistent ethical theory based on equality between all sentient beings.

Chapter 9 The sentience problem

I already mentioned that there is a gradation in mental capacities. Some beings have a richer, more complex emotional life than others. But there is also the scientific question: which beings are sentient? It is a scientific question, because being sentient is a matter of fact, and science is about discovering facts. Science can determine criteria for sentience. The sentience problem then consists in determining those criteria, and testing animals to see whether they satisfy those criteria. The latter raises real ethical concerns. Below I discuss the scientific and ethical problems of sentience.

9.1 The scientific problem

There are four criteria to see whether a living being is sentient.

1) *The adaptive role of feelings.* Pain and other feelings can offer an evolutionary advantage to living beings. Feelings of pain can result in avoiding some behavior, learning, protection or self care. But we have to be aware that the organ that generates feelings (the brain), might consume a lot of energy. So not every living being will invest in such an organ. For living beings who cannot move in complex ways, having feelings is useless. Plants cannot show the fight, flight or freeze responses, so fear is not useful for them.

2) *Anatomical basis.* We know that pain is related to specific neurons (nociceptors or pain receptors) and other anatomical properties (e.g. the central nervous system). Also other feelings, such as fear, are related to specific parts in the central nervous system (e.g. the limbic system and the amygdala).

3) *Behavior.* Feelings are often associated with specific behavior (e.g. the fight, flight or freeze response of fear). Healing wounds, scratching, loss of sexual interests, vocalizations, body movements, facial expressions and many other things might indicate that the living being feels pain. Especially when the behavior persists for a long

time, is repeated quite some time after the event, or changes in more complex ways, we can exclude the possibility that the behavior is just an automatic, instinctive or reflexive response.

4) *Physiology*. When there are other changes in the body (e.g. faster respiration rate, heartbeat, blood pressure, eye movements,...) then we might expect emotions to be present. Also the effect of some chemicals (endomorphines and analgesics) on the behavior might indicate sentience.

With these four indicators, we can test whether an animal is sentient. According to the current scientific consensus (Masson, 1995; Griffin, 2001; Bekoff, 2007) most likely vertebrate animals with a functioning central nervous system are sentient and can feel pain, fear and distress. Perhaps some squids and large crustaceans are also able to subjectively feel something.

Let's take a concrete group of vertebrate animals: fish. The following citations represent the current scientific consensus (EFSA, 2009):

"There is scientific evidence to support the assumption that some fish species have brain structures potentially capable of experiencing pain and fear. The balance of evidence indicates that some fish species have the capacity to experience pain. [...] Responses of fish, of some species and under certain situations, suggest that they are able to experience fear. [...] From studies of sensory systems, brain structure and functionality, pain, fear and distress there is some evidence for the neural components of sentience in some species of fish. Our knowledge and understanding of manifestations of sentience in fish, however, are limited. [...] From studies of sensory systems, brain structure and functionality, pain, fear and distress there is some evidence for the neural components of sentience in some species of fish. Our knowledge and understanding of manifestations of sentience in fish, however, are limited. [...] The stress physiology in fish is directly comparable to that of higher vertebrates."

We see that the statements are very prudent, because there are so many fish species, and only a few are studied (the most famous studied species are rainbow trout). The fish that were studied indicated signs of sentience. The following is a list of 11 criteria to test the presence of a pain system.

1. There should be pain receptor cells present.
2. There should be a nociceptive neural pathway from the tissue to a higher brain structure.
3. In this brain structure, there should be specialized processing systems that are active when the tissue is damaged.
4. There should be specialized transmitter substances along the neural pathway.
5. In the specialized brain part, there should be endogenous opioids and opioid receptors.
6. There should be electrophysiological responses to cuts and bruises.

7. There should be a suspension of normal activity associated with noxious stimuli. For example eating and sexual activity should stop.
8. There should be behavioral change and avoidance in the short term (moving body parts, scratching, avoiding the threat).
9. There should be learned avoidance of places in the long term.
10. There should be a measurable influence of analgesics in reducing responses.
11. There can be effects of chronic stress (e.g. a malfunctioning immune system after long exposures to pain).

All these criteria are fulfilled for a rainbow trout. Trout have nociceptors and a neural pathway from these pain receptors to their brains. Scientists have injected a bee venom in the lip of a trout. The trout stopped eating, and refused to eat for quite some time. He started scratching his lip in the sand. The trout showed less avoidance when new fearful objects were placed in the water, as if the trout was so concerned with his pain that he became less aware of his environment. But after the trout was injected with some analgesics, he stopped scratching, became more aware of his environment and swam away from fearful objects (Sneddon et al. 2003a; 2003b).

Consistency in judgments implies that we have to give those fish at least a strong benefit of the doubt. If a mentally disabled human (who was not able to talk) showed similar reactions in similar experiments, we would judge this human to be sentient. Therefore, we should have the same judgments towards those fish.

When it comes to invertebrates, things get more difficult, because not all of the above criteria are satisfied. Some insects simply continue eating after they lose a leg. Sentient beings would likely stop eating, no matter how hungry they are. We could give insects the benefit of the doubt, by not eating them.

Plants are even less likely to be sentient, as none of the criteria are fulfilled. Some plants are able to react to a threat (e.g. produce poisons to protect themselves against herbivores), warn other neighboring plants, and communicate in rather sophisticated ways. Some plants might even have a mechanism for self-recognition (Karban, 2009). Although this is nothing yet compared to (self)consciousness, those plants have roots that can recognize whether other roots belong to the same plant or not. These are enough reasons to grant plants also some basic right, as we have seen in the chapter on the basic right (section 6.4). But we should not yet conclude that those plants are sentient. Our immune system and computers also have very complex patterns of communication and self-recognition, but that does not yet make them conscious systems. It might be possible that plants have a yet unknown system that makes them sentient. But such a system is not yet found, and for the same matter, it might be possible that the hairs on my head are sentient beings. Perhaps they also have a special system?

If we would give plants the benefit of the doubt, placing them at the same level as real sentient animals, then we have to realize that we have to eat plants in order to

survive. Eating plants is a survival end, and if plants become equal to all other sentient beings, then we are also allowed to eat beings who are definitely sentient. We cannot run that risk based on the current very weak evidence, so we need much more evidence before we should say that plants are sentient.

Although we do not know yet how sentience is generated by the brains, I believe that at the most basic level, having an affective feeling is binary: either you have it or you don't. Once a being has at least one affective feeling that reflects a need, it is sentient (with respect to that need) and it has a well-being. The above 11 criteria are nothing but rules of thumb to determine from the outside whether someone experiences pain.

The reason why I believe in the binary nature of sentience is that I cannot imagine myself to have a half feeling. Either there is light on the stage (even if it is a dim light), or there isn't. This distinguishes sentience from other mental capacities such as rationality. Rationality is more difficult to define, and it has less this on or off characteristic. There are multiple aspects behind rationality (you need e.g. self-consciousness, some imagination and a memory), and sometimes not all aspects are present.

Compare it with letters on a paper. Being sentience is like having a stain of ink on a paper: either there is a drop of ink on the paper (even if it is a tiny drop), or there isn't. Rationality is like having the letter A on a paper: this requires much more aspects. It is not always easy to see whether the ink stains form a letter A or not. From a scientific perspective, it is in theory possible to see the presence of ink stains: one only needs a tool to see the stains (e.g. a good microscope). If such a tool does not yet exist, it does not mean that the presence of ink stains is not a matter of fact. However, determining whether the stains form a letter A always involves an element of (subjective) interpretation.¹

9.2 The ethical problem

We have seen criteria to determine whether a being is sentient, whether it can experience pain, fear or distress. I mentioned an experiment that caused pain to

¹ The analogy between mental states and ink stains can be explored further. For example if ink stains are connected, they belong to the same pattern (e.g. the same letter), just as different feelings can be connected to belong to a same person. There can be fuzzy boundaries between ink stains, just as there can be fuzzy boundaries between mental states, generating the problem of psychological connectedness explored in section 4.2.4.

rainbow trout. The problem is: animal experiments are often violations of the basic right, because the being is used as merely a means (experimental object). So that would mean we would not be allowed to do tests to determine whether an animal is able to feel. That means we could not so easily determine whether the animal is able to feel. Perhaps we might never know, even if it would be scientifically possible to know.

So, are those tests morally permissible? We can argue that they are, because we could say that we do not really violate the basic right of such an animal when our goal is to determine whether it has the basic right. In other words: the animal is used as merely a means to an end, and the end is determining whether the animal has a basic right. This end is not a survival end, neither is it a vital need, a basic need or a luxury need. The end is of a totally different moral category.

Actually, this idea is not farfetched. We already have a similar approach towards humans. What do physicians do when they have a patient who does not seem to react to certain impulses? Sometimes the physician tests the patient, to see whether she still reacts to painful stimuli. This is a test that might cause pain, very similar to the tests with the rainbow trout. It is not disrespectful towards the patient.

Chapter 10 The predation problem¹

In this chapter we arrive at the last big problem of animal ethics. It is a very serious one, as it might punch a big hole in a consistent animal rights ethics. The problem receives some attention from time to time (e.g. Cohen & Regan, 2001; Cowen, 2003; Ebert & Machan, 2012; Everett, 2001; Fink, 2005; Horta, 2010; Sapontzis, 1984; Simmons, 2009), but the challenge that this problem poses is however underestimated by both animal rights ethicists and critics. It is strange why animal rights critics do not toss this problem about more regularly in discussions, if their goal is to show that an animal ethics is inconsistent. To present the problem as clearly as possible, let's start with the following two scenarios.

1) *The predation dilemma*. A lioness is going to attack a zebra in order to feed her two hungry whelps. You are sitting in a car, looking at the scene. You can easily save the zebra, simply by turning on the engine of your car, chasing the lioness away. Should you save the zebra?

2) *The transplantation dilemma*. In a hospital two children need new organs, but no organs are available. A surgeon is about to kill a visitor against his will, in order to use his organs to save the two patients. Is he allowed to do so, or should we interfere and stop him?

A lot of people, including most animal rights activists², say that we do not have a duty to protect zebra from lions, but we do have a duty to protect the visitor from the

¹ Sections 10.3, 10.4, 10.6 and 10.7 are based on *Overcoming Problems in Consequentialist Welfare Ethics. On predation, motion, organ transplantation, medical experimentation and procreation*, Bruers & Van den Berg, submitted to *Utilitas*. Sections 10.4 and 10.5 are based on *Moral Intuitions and the Intrinsic Value of Biodiversity*, submitted to *Environmental Values*.

² This claim is based on a personal (unpublished) survey that I did with more than 30 animal rights activists. 90% of them would not condone organ (xeno)transplantation, whereas only 7% would not condone animals preying on animals and 14% would not condone animals preying on humans. Also, another 24% of respondents were undecided in the latter case. The preying on animals and organ (xeno)transplantation dilemmas showed only 0-6% undecided responses. This demonstrates that the problem of animals preying on humans is the most difficult dilemma for animal rights activists, and that in the two predation dilemmas (preying on animals and on humans), animal rights activists are more speciesist than in the two organ (xeno)transplantation

surgeon. The killing of the visitor against his will should be prohibited. Now, if we adopt an antispeciesist ethic, we have to be able to switch the positions between animals and humans. Hence, what if, instead of a zebra, there was a lonely human child being attacked? Or what if instead of killing a visitor, the surgeon would kill a non-human animal (e.g. a pig) for xenotransplantation (and suppose that xenotransplantation works)?

For an ethic based on animal equality, there are in fact two problems of predation and transplantation.

1) What is the morally relevant difference between predation and transplantation? After all, in both cases (meat consumption and organ transplantation) a sentient being (lioness and surgeon) kills another sentient being (zebra, human or pig) without permission and uses parts of its body (muscle tissue and organ tissue) by taking these body parts up in the bodies of vulnerable sentient beings (whelps and child patients) in order for them to survive. The analogy could not be clearer. Yet, most people, including defenders of animal rights, condone the predation but condemn the transplantation (antispeciesists also condemn xenotransplantation). The problem arises because our moral intuitions say that in one life-and-death-dilemma action is permissible, whereas in the other dilemma it is not.

2) In the predation situation, the animal rights activist would protect the human child, but most activists would not protect the zebra. Does this mean they are still speciesist?

These two problems (the morally relevant difference between predation and transplantation, and the difference between a zebra and a human as prey) will be called the *difference problem* and the *prey problem* respectively. In my opinion, these two problems combined are the weakest spot in a consistent antispeciesist animal rights ethic. They are underestimated by both animal rights advocates and critics.

The predation problem can be a strong weapon in the hands of a speciesist (such as Cohen in Cohen & Regan, 2001, p30), because s/he could easily solve both problems with one stroke, by claiming that there is a morally relevant difference between humans and non-human animals. According to speciesism, all individuals belonging to the human species have a higher moral status than everything else. In that case, everyone is allowed to use animals for xenotransplantation and predation. But lions are not allowed to use humans. We have a duty to protect humans, but no duty to protect zebras. By solving both problems with this one simple criterion, it seems that the speciesist ethic is coherent. But is it?

dilemmas (using either humans or pigs as organ donors). However, the order of presentation of the dilemmas also influences the judgments.

As we have seen in a previous chapter, the species boundary cannot be morally relevant. The five arguments against this criterion show that the speciesist solution is not coherent after all. It is at least as arbitrary and artificial.

What I will do in this section is present solutions to both the difference and the prey problems that are consistent with an antispeciesist theory of animal equality. Doing this we can avoid arbitrary elements in our ethic, and the resulting ethic fits with our strongest moral intuitions.

The principle of tolerated choice equality might solve the prey problem, whereas the triple-N-principle (referring to three criteria: normal, natural and necessary) and a 'fairness' principle might make a morally relevant distinction between predation and transplantation. I demonstrate that the triple-N-principle is in agreement with moral intuitions that a lot of people (both speciesists and animal rights activists) have, and that it is in correspondence with the moral value of biodiversity. The tolerated choice, fairness and triple-N principles might make our antispeciesist ethic much more coherent than the speciesist one.

10.1 Invalid solutions to the prey problem

Starting with the prey problem, I first briefly mention a few invalid solutions, given by some animal rights ethicists.

Some people claim that humans are not part of the natural diet of lions (e.g. Simmons, 2009). So lions are only allowed to eat what is part of their natural diet. One problem with this argument is that it poses a strange, arbitrary distinction between humans and non-humans. Isn't it strange that of all species on earth only *Homo sapiens* never are or have been natural prey? More importantly: when is something part of someone's natural diet? If a lion wants to hunt a human, is the human not part of his preferences, and hence part of his diet? As rabbits were introduced in Tasmania some 200 years ago, they can be considered as not being part of the natural diet of a Tasmanian devil. So we now have a duty to protect rabbits from Tasmanian devils?

Another claim is that "Duties of assistance exist only insofar as potential beneficiaries require assistance in order to flourish according to their nature" (Everett, 2001 p55) It is believed that we don't have a duty of assistance towards zebra, because they can flourish without our assistance. On the other hand, children cannot flourish without our assistance, so we have a duty to protect children. The problem with this approach is that it seems to make an arbitrary distinction between flourishing according to one's nature or not. What if the zebra is injured? Doesn't he require assistance then? And what does

“flourish according to their nature” really mean? It resembles some kind of essentialist thinking that we countered in a previous chapter.

10.2 A hypothetical solution to the prey problem

When asking animal rights activists whether they would save the human, they said they would, although they would not necessarily save the zebra. They had real difficulties with those scenarios and were tempted to be what they thought was speciesist, and consequently claimed that true antispeciesism was impossible to achieve.

The solution to the prey problem that I would propose, is based on tolerated partiality (or tolerated choice equality). So the prey problem is a situation where this equality comes in handy. It is a principle that helps us make our moral intuitions (at least a bit more) compatible with antispeciesism.

As in the burning house dilemma, we are allowed to save the sentient beings with whom we feel a strong connection or empathy. Thus we should tolerate a choice to protect the lion instead of the zebra or to protect the child instead of the lion. We have the right to be partial, as long as we respect equal partiality from others. If someone saves the human, s/he should tolerate the choice of someone else saving the zebra instead of the human. The tolerated choice equality allows us to reconcile two preferences of the antispeciesist: on the one hand s/he would (in most cases) save a human rather than an animal, but on the other hand s/he doesn't want to be speciesist. We are not speciesist if we tolerate the choices of someone who protects the zebra or the lion (i.e. someone who did not protect the human). Everyone is free to choose whether to protect the zebra, the human or the lion.

Empathy is crucial in the tolerated choice equality; it is the driving force to help vulnerable beings. If someone feels more empathy and connection with one being than with another, this justifies his preference for saving the preferred being. We should tolerate his/her choice, because empathy is a moral virtue and s/he acted with empathy as long as s/he did not hate or disdain the other being.

For speciesists this tolerance of saving the zebra or the lion instead of the human is likely the hardest nut to crack in the whole theory of animal rights. It might be in rather strong contradiction with one of their moral intuitions. In practice however, they should not be so concerned, because asking what they would do, most animal rights activists responded that they would save the human anyway.

There is one more thing. As we have seen in the chapter about the basic right, we could introduce a criterion that refers to a higher moral status related to some mental capacities like self-consciousness, moral agency or rationality. So we might have a

stronger duty to protect those sentient beings who possess those special mental capacities, or who will develop them, or who have close relatives (e.g. parents) with such special mental capacities. The only thing a speciesist then has to accept is tolerated choice equality between seriously mentally disabled human orphans and non-human animals such as zebra. If we say we have a duty to protect those disabled orphans, whereas we do not have a duty to protect non-human animals because all humans have a higher moral status than non-humans, then we become too partial. It is a kind of speciesism, and like racism or sexism it is a kind of partiality that we cannot tolerate.

In summary, we do not have a duty to defend the prey from predators, but we are allowed to defend the prey if we feel an emotional need to do so. (We can add that we are only allowed to defend the prey if this does not result in severe ecological damage.) In this permission to defend the prey, we have a right to be partial to some degree (not too much), as long as we respect similar levels of partiality of others. And if the prey has some special mental capacities, we might have a stronger duty to protect it.

10.3 Invalid solutions to the difference problem

Next, we move to the difference problem. Again, I first briefly discuss a few invalid proposals encountered in the literature to solve the problem of the difference between predation and organ transplantation.

1) Feasibility: stopping transplantations is feasible, stopping predation is not. Peter Singer gave a practical argument: It is impossible to intervene in nature to protect all prey animals (Singer, 1990, p226). However, a single intervention, such as saving a zebra in front of you, is feasible. The feasibility argument is related to the demandingness objection against consequentialist welfare ethics, which can be shown to be impotent in some way (Sobel, 2007).

2) Moral agency: surgeons are moral agents, lions are not. Regan's answer (Regan, 1983) is that carnivorous animals (e.g. lions) don't have moral reasoning capacities; they are not moral agents. As amoral beings, they don't have duties of non-maleficence and beneficence. Furthermore, according to this view, we as moral beings do not have a duty of beneficence towards a victim when the agent (aggressor) is an amoral being.

The next table presents the duties of interference that we have when a moral or amoral agent (aggressor) attacks a moral or amoral victim.³

		Victim	
		Moral being	Amoral being
Agent	Moral being	Obligation	Obligation
	Amoral being	No obligation	No obligation

The table shows that we only have obligations of beneficence towards the victim when the aggressor is a moral being. But this moral agency account faces some counter-intuitive problems. First, what if amoral predators attack moral humans? Regan is not clear about our duties of saving those humans, but even he points at an intuition that at least it cannot be wrong to interfere and save the human.⁴ Second, a lot of animal rights ethicists believe that we do have a duty to protect sentient beings from amoral threats such as falling rocks. Third, what if some animals were moral or gain moral consciousness? At this moment no non-human animals are real moral agents, although dolphins and great apes might come close: they have a self-consciousness, a high (social) intelligence and premoral sentiments (Shermer, 2004, p. 16). And dolphins kill and eat sentient fish. A strong ethical theory should be able to deal with counterfactual situations: what if dolphins gained moral agency? Do we then have a duty to interfere and save the fish? Fourth, what if some humans needed meat to survive? My intuition tells me that we do not have a duty to stop moral humans or moral dolphins from hunting if those humans and dolphins need meat in order to survive. After all, it seems strange why non-rational (amoral) predators should have an unfair advantage against rational (moral) predators: the former can hunt, the latter can't. (For further criticism of Regan's argument, see also Jamieson, 1990 and Fox, 1999 p.163.)

3) Libertarianism. Related to the moral agency account, is the theory of libertarianism (Ebert & Machan, 2012). This theory says that moral agents have a duty not to harm others, but they do not have a duty to protect others (e.g. protect prey). In

³ Suppose that interference (beneficence towards the victim) always harms the agent. For example interfering in predation harms the predator by limiting its food supply.

⁴ As discussed in Ebert & Machan (2012), Regan's theory would imply that it is prohibited to save the human prey. But elsewhere, Regan claimed that 'it cannot be wrong to do what will harm the child [who has come into possession of a loaded revolver and has begun to fire it at us], even though the child is innocent and so does no wrong.' (Regan 1983 p. 293). It is not clear why our obligation to interfere in this situation with the armed child would not extend to the situation of the predator.

other words: amoral beings do not have rights of beneficence, a right to be protected from harm. So libertarians say that we do not have a duty of beneficence towards a victim when the victim is an amoral being. But all moral agents still have a duty of non-maleficence towards moral and amoral sentient beings. The next table presents our duties of interference to help victims from aggressors (agents). This libertarian theory is in some way a ‘transposition’ of Regan’s account.

		Victim	
		Moral being	Amoral being
Agent	Moral being	Obligation	No obligation
	Amoral being	Obligation	No obligation

This libertarian theory is also plagued with some problems. First, mentally disabled humans are amoral beings and hence have no right to be protected.⁵ This seems counter-intuitive to me. Second, hypothetical carnivorous moral agents (e.g. moral dolphins) would have a duty of non-maleficence, so they should not hunt. Although we do not have an obligation to stop the dolphins from hunting, the moral dolphins themselves should now abide the rights of sentient fish. Similarly, moral humans have a duty not to move when insects are sentient (because moving around harms those sentient insects and animal rights libertarians say that moral agents have a duty of non-maleficence towards all sentient beings).

4) Group protection: lions form a group, patients in the hospital don’t. Extinction of a group is worse than the death of a number of single individuals. But first, a group (e.g. a species) is an abstract set of individuals. Does an abstract set in itself have interests? Why give moral status to this abstract set? Second, a group is an arbitrary set. What group should we take? Should we consider the species of lions (*Panthera leo*) as a group? Why not take a specific population of lions as a group? Or for that matter the genus of *Panthera*, the family of felidae, the order of carnivora, the class of mammalian

⁵ Ebert & Machan discussed an extension of this strict libertarian ethic, where we have special duties towards some amoral beings, based on special relationships we have with them (e.g. parental relationships). But is the relationship that I have with a mentally disabled *orphan* more special than the relations that I have with amoral animals? The problems with this extension is: 1) what counts as a sufficiently special relationship in order for us to have a special duty? And 2) how to avoid arbitrariness (e.g. reference to arbitrary biological classifications such as species) in this extended libertarian theory? A non-arbitrary extension that includes the rights of mentally disabled orphans to be saved from harm, results in a complete animal rights ethic, where we have a duty to help amoral animals as well.

or the phylum of chordata? And if we have siblings (or twins) in the hospital who both need organs, aren't these siblings a group in some sense? They have some unique properties in common. Third, the group argument is artificial. Consider the lonely lion problem: what if there is only one lion left in nature? Is this one lion still a group, or a single individual? Is he allowed to hunt? Does the right to hunt depend on the presence of other similar individuals?

5) Illness: the patients in the hospital are ill, the zebras aren't. But what does 'being ill' mean? Both the patients and the whelps are feeling very sick when they don't get something. The symptoms might be quite similar. Both will die eventually. For a predator, meat can be considered as a medicine to stay healthy. Second, this criterion doesn't seem to be quite empathic towards the ill persons. Why should you lose the right to use someone (violate his rights) if you are ill, and are you allowed to use someone if you are not ill?

6) Existence: lions would not even exist if they were not allowed to hunt; surgeons and patients in the hospital were born, even if forced transplantation was not allowed. But suppose that we have a boy who has a genetic defect, and this boy needs a new organ in order to survive. Also, the mother of this boy had the same genetic defect, and 30 years ago they did a successful organ transplantation to save this woman. The woman survived, and gave birth to a son with a similar genetic defect and thus a similar disease. In this case the boy's very existence depends on the transplantation for his mother. Should we now make an exception and tolerate the coerced sacrifice of someone for organ transplantation to save this boy?

7) Ecological disasters: preventing predation would result in ecological catastrophes, whereas prohibiting transplantation would not. This argument actually makes some sense.⁶ If we would consistently intervene in nature to prevent predation, then a lot of predator populations might die of starvation, which would have ecological effects on prey species (e.g. overpopulation, increased competition and spread of diseases). It is very difficult to calculate the overall effects on animal death and suffering, because ecological interactions can be very complex. From a precautionary principle we might say that at this moment it is better not to consistently intervene in predator-prey interactions in nature. This means we do not have a duty to intervene. Of course, once (in a far-away future) ecologists would be able to calculate that the extinction of predator species would be good on the whole, then we should go for predator extinction.

⁶ This point was already made by Singer in 1973: "Lions play a role in the ecology of their habitat, and we cannot be sure what the long-term consequences would be if we were to prevent them from killing gazelles". Many other philosophers made this point (e.g. Simmons, 2009).

8) Uncertainty aversion. This is related to the previous proposal of ecological disasters. It can be derived from the thought experiment of impartiality (Rawls' veil of ignorance), as described in the section about prioritarian ethics. We have seen that from behind the veil of ignorance, one's level of risk aversion was important, resulting in a theory of quasi-maximin prioritarianism. I also mentioned that next to risk aversion, uncertainty (or ambiguity) aversion is important as well. We have seen that the solution of the trolley problem (pushing a heavy man from a bridge) could depend on one's uncertainty aversion behind the veil of ignorance. To explain uncertainty aversion, I presented Ellsberg's paradox (Ellsberg, 1961), and made a connection between this paradox and the trolley problem. What I will demonstrate now, is that the very same idea of Ellsberg's paradox might solve the difference problem.

The following version of Ellsberg's paradox will be useful in this discussion. An urn contains three balls of two different colors: green and red. You win the game when you draw the green ball. You can choose between two games of chance. In the first game, you know that one ball is green and the others are red. Hence, your probability to win is exactly $1/3$. In the second game, you only know that at least one ball is red. Now your probability to win is between 0 (when all balls happen to be red) and $2/3$ (when the two unknown balls happen to be green). Which of these two games do you prefer to play? If you have – like many people – uncertainty aversion, you'd prefer to play the first game, because in that game you at least know your chances to win.

We can simplify the predation problem to see the analogy with the above Ellsberg paradox. Suppose from behind the veil of ignorance you know that you will be born as one of three sentient beings: one predator or two prey animals. You now that the predator needs two prey in order to survive. You can now decide whether predation is allowed or not. You have to choose between two games of chance. In the first game, a world with predation, you know that the predator is going to kill and eat the two prey. These two prey lose, they are the red balls, and red means dead. You have $1/3$ chance to be born as this predator and survive. So your chance to win is $1/3$. In the second game, a world without predation, you know that the predator will die from starvation (he is the red ball). But what happens with the two prey? If there are enough resources, they can both survive (they are both green balls). But likely there are not enough resources for both. So they might start to fight until one or both of them die. Or they might overexploit the resources so that one or both of them eventually die from starvation. You don't know what will happen, and most of all: you don't know the probabilities for them to survive. Your chance to win is something between $2/3$ (if there are enough resources for both prey) and 0 (if everyone dies). Uncertainty aversion implies that you would prefer the world with predation.

The difference with the organ transplantation goes as follows. There are two patients and one visitor. You can be one of them. If you choose a world where forced organ transplantation is not allowed, you will win when you are the visitor, because he will

survive (he is the green ball). However, if you choose a world with transplantation, the visitor is sacrificed against his will, so he will be the red ball.

If the bodies of both patients accept the organs, they can both survive. But if one or two of the transplantations are not successful, one or two of the patients die. You do not know the success rate, so your probability to win will be between 0 (everyone dies) and $2/3$ (both patients survive). Uncertainty aversion implies that you would prefer the world without transplantation.

A lot of people have uncertainty aversion, and this might already justify the choice of a world with predation and without transplantation. Uncertainty aversion is able to make a difference between predation and transplantation. But even if we respect uncertainty aversion as a cognitive bias, this uncertainty aversion account still faces some problems. First, it implies that moral duties depend on our current, contingent, subjective state of knowledge. If knowledge about ecosystem functioning increases, we might need to change our judgments about predation, and our duties towards prey.

Second, in a lot of situations, the uncertainty aversion gets 'skewed'. Consider a more realistic predation problem behind a veil of ignorance. Suppose one lion eats hundred zebras in order to stay alive. So there are 101 individuals: one lion and hundred zebras. You can be born as any of those animals. You can again choose between two worlds. In a world with predation, the lion eats all the zebras, so you have probability $1/101$ to be born as the lion and survive. In the world without predation, the lion definitely dies, but as a consequence of increased competition and ecological overshoot, some of the hundred zebras might die or start fighting for scarce resources and kill each other. You now have an uncertain probability between 0 and $100/101$ to be born as a zebra that is able to stay alive. It's a game of chance, and you have to choose between a certain probability $1/101$ versus an uncertain probability between 0 and $100/101$ to win (to survive). Even if the second game has uncertainty, a lot of people would still prefer to play that game, because $1/101$ is close to 0 and hence the uncertainty is 'skewed'.

Third, there might be people who think that even without uncertainty, we still do not have a duty to interfere in predation. If it seems counter-intuitive that our attitude towards predation, and the fate of predators, depends on our subjective state of knowledge, we have to look for a more principle-based ethic that allows predation. So in the next section I want to explore a justification of predation that might satisfy the needs of those people. The principle that I propose can be expressed as a 3-N-principle, related to the intrinsic value of biodiversity.

10.4 A first hypothetical solution to the difference problem: the 3-N-principle

When it comes to justification for predation, a lot of people also give answers that are typically used by meat eaters: they say it is ‘necessary’, ‘normal’ or ‘natural’ for the lion to hunt. Note that these are the three N's that Melanie Joy pointed at in her discussion of carnism, the often hidden ideology of most meat eaters in our culture (Joy, 2001; 2009). So even some animal rights activists are still tempted to use the same ‘naturalistic arguments’ that meat eaters (carnists) use.

What I am going to do now is 1) clarify the criteria ‘necessary’, ‘natural’ and ‘normal’ to make the principle more accurate, 2) demonstrate that none of those three criteria are separately valid (none are sufficient to violate rights), 3) show that the combination of all three criteria might be in line with our moral intuition (predation is natural, normal *and* necessary and therefore allowed; transplantation is not), 4) argue that there is a connection between the three N-conditions and biodiversity, and 5) argue why biodiversity can be morally relevant to allow rights to be violated. If this strategy works, we not only have a consistent animal ethic that is compatible with the moral intuitions of most activists and other people (e.g. with the intuition that we should not intervene in predation), we also have a principle which is in essence based on criteria that carnists already use. So the speciesist/carnist would have more difficulties in replying that this 3-N-principle is not a good solution.

First, we have to clarify the three criteria. The condition ‘necessary’ indicates a sufficiently strong vital need for an individual (e.g. food) or a group of individuals (e.g. procreation). ‘Normal’ means for simplicity that something occurs often. If apples often fall from trees, it is normal. The criterion ‘natural’ points at everything that is created by evolution. Evolution is the aimless process of genetic mutation and natural selection. Conscious, intentional inventions by intelligent beings are not considered natural, although the intelligence of those beings is natural (arose by evolution). A process is natural if it originates from natural evolution instead of being created by artificial means (such as conscious inventions).

Second, I argue that none of the three N criteria separately are sufficient conditions. Necessity is not a sufficient criterion: for the patient in the hospital it is necessary to get a new organ in order to survive. So necessity alone does not distinguish between predation and transplantation. Yet, in ethics, it seems intuitively clear that necessity has some moral significance.

What about natural? Quite often a carnist argues that eating meat is natural for humans, and therefore it is allowed. But there is some danger in applying the criterion of naturalness. Is rape natural? After all: by natural processes (evolution) men have

developed an body part that makes them capable of raping women, quite a lot of male animals (e.g. cats) rape female animals in nature, also humans rape each other, our ancestors most likely raped each other for thousands of years, and it is even very likely that we owe our very existence to the fact that some of our ancestors once raped a woman. So what else do we need to say that something is natural? Simply put: rape is natural (and normal in the animal world), but not necessary for men. Therefore it is an unnecessary violation of rights, and hence immoral, even if rape was as natural as one could think of.

One could say that the patients in the hospital got ill by natural processes. But this does not make organ transplantations natural. Transplantation was a conscious invention. It did not evolve by a blind process of mutation and natural selection, it is definitely not instinctive behavior.

The ancestors of carnivores on the other hand did not consciously think about a problem of nutrient shortage, intentionally look for a solution, experiment a bit, discover meat as a solution, whereupon they adapted and suddenly became dependent on meat. Predation is therefore natural because it originated by a process of evolution (mutation and natural selection).

For carnivores and omnivores (like humans), we could say that meat consumption is natural, because we have developed - by blind evolution - a digestive system that makes it possible to eat meat. Eating meat also often happens instinctively, which is not the case for transplantations. But for humans, meat consumption is, just like rape, not necessary (ADA, 2009). As it is an unnecessary violation of rights, it should not be allowed, even if it was as natural as one could think of. We should eat vegan instead. And finally, note that intensive livestock farming is a far cry from naturalness. So animal products from farming are neither necessary nor natural.

And what about normality? If rape happens a lot, that does not yet justify rape. If killing happens a lot in wartime, it does not justify murder.

All three criteria considered separately are not valid as moral arguments. Also combinations of two of the three criteria are not sufficient. If rape is natural and happens quite a lot, it is still immoral because it is not necessary. If organ transplantation is necessary and happens quite a lot in some distant country, it is still immoral because it is not natural. And if predation is natural and necessary for the predator, but not normal, then I guess we are tempted to stop this predation. If every being on earth was vegan and then suddenly by a process of natural selection a small group of animals appeared who needed to kill and eat other animals in order to survive, animal rights activists would likely intervene (if they could). They would not allow those predators to kill others.

So let's put all three criteria together: suppose something is normal, natural *and* necessary. What if rape was not only natural, but also normal and necessary for humans? In other words: what if all human populations would go extinct if they were

not allowed to rape? Extinction on a massive scale. Then that's some argument. Although I don't know what most people would say, we might tolerate rape in that case.

I believe that predation by natural predators is allowed because it is normal, natural and necessary, and transplantation is not allowed because it is not normal and will never be natural. The 3-N-conditions put together might solve the difference problem. So let's now try to formulate the above normal+natural+necessary criterion in a more exact and complete 3-N-principle:

The 3-N-principle. If (a) a sufficiently large group of sentient beings became by (b) an evolutionary process (c) dependent on the violations of rights of other sentient beings for their survival, they are allowed to violate those rights for that purpose. (But we are also allowed to protect ourselves and to protect prey if we are inclined to do so, and we have a duty to intervene in predation once feasible alternatives such as healthy non-animal food for the predators are found. In practice, we should give dogs vegan food.)

This principle clearly refers to (a) normality, (b) naturalness and (c) necessity. Of course there can be fuzzy boundaries between normal and not-normal, natural and not-natural, necessary and not-necessary. Some things can be very normal, natural and necessary, others less so. The idea is that these fuzzy boundaries create a gradation and that this gradation can be coupled with the gradation of rights violations: some actions are strong violations, others less so. An example of such a coupling of gradations was presented in Figure 9, section 6.4).

10.5 The value of biodiversity

The above 3-N-principle makes a distinction between predation and transplantation, and this is in line with the moral intuitions. But what is the moral relevance of those three criteria? Consider the no-harm principle. This principle corresponds with our moral intuitions, but there is more: there exists a natural property called well-being. If we give intrinsic value to this natural property, then the intuitions behind the no-harm principle are coherent with this value of well-being.

Could we do the same for the 3-N-principle? Does there exist a natural property that we can value and that is coherent with the 3-N-principle? The answer is affirmative: I suggest that the 3-N-principle is connected to the moral relevance of biodiversity. Let's

define biodiversity as everything that directly originated from evolution. If now we would suppose that biodiversity is morally very important, then naturalness is relevant.⁷ Or in other words: if a process (behavior or property) is natural, it contributes to biodiversity by definition. If a process is natural and normal, it contributes a lot to biodiversity. And if a process is natural, normal and necessary, biodiversity would drastically decrease if that process no longer exists. A drastic decrease of biodiversity can be considered worse than violations of rights, and this makes the connection between the 3-N-principle and the value of biodiversity. The connection is only valid if all three N-criteria are present. Applied to predation: the existence of all predators in the world strongly contributes to biodiversity, and this biodiversity from predation has a moral value that strongly trumps the basic rights of individuals (e.g. the right not to be used as merely means), and also trumps utilitarian calculations of well-being.⁸

There is one more assumption required to make the above connection tight: duties and moral principles should be universalized (compare with the universal law

⁷ This is similar to Goodin's "green theory of value" (Goodin, 1992). Goodin claimed that things created by natural processes possess a higher value than things created by artificial processes. The intrinsic value of biodiversity plays a central role in many holistic environmental ethics (Benson, 2001; Rolston, 1988).

⁸ There is a subtle issue relating the three N criteria to biodiversity. Do acts (such as the behavior of predation) itself contribute to biodiversity? Or is it rather the existence of predators that contributes to biodiversity (even if the predators manage to survive on vegan food and stop predation)? In other words: is biodiversity the variation in different entities (life forms, genes,...) or does it include variation in processes? If processes are not included, the 3-N-principle and biodiversity principle are slightly divergent, because the 3-N-principle refers to processes (types of behavior). I tend to think if processes contribute to biodiversity, they only do so slightly. As a result, a world where lions live but survive on vegan food is better than a world where they live and hunt prey. The suffering of the prey is worse than the loss of process-biodiversity (the loss of the behavior of predation).

Another potential difference between the 3-N-principle and the biodiversity principle is that the necessity condition in the 3-N-principle might be extended to include not only a vital need but also a need for sufficient health. Imagine that we do not need meat for our survival, but that meat would positively contribute to our health (i.e. a vegan diet would be less healthy). If that were the case, eating some meat might be permissible according to the 3-N-principle if necessity includes health, but it might be impermissible according to the biodiversity principle, as refraining from consuming meat does not result in the extinction of humans. If we want a consistent ethical system where eating meat would be permissible when veganism were unhealthy (but not deadly), we should stick to the 3-N-principle instead of the biodiversity principle. Some animal rights activists would eat vegan, even if it had some costs to their health. A related difference between the necessity and biodiversity criteria lies in the answer to the question: should predators stop eating once they have procreated? As soon as predators have procreated and have viable offspring, the predator population can survive. One could say that biodiversity does not decrease when all predator parents die. Hence, when a predator has offspring, meat is still necessary for the live of this individual, but not for the conservation of biodiversity. On the other hand, one could say that the existence of predator parents does contribute to biodiversity, as if the group of predator parents form a separate population. The number of predators will drastically increase when all predator parents die. And this decrease can be counted as a drastic decrease of biodiversity.

formulation of the categorical imperative of Kant). We do not have a duty to protect the prey, even if we could in a particular situation, because if we did say we have a duty in this situation, then we should want this rule to be universalized to all predators. Then we should want predation to be prohibited at all times, anywhere.⁹ But predation is necessary for predators, so this universalization implies that we should want all predators to go extinct. Predation is also normal, so a lot of predators would go extinct after universal prohibition. And that would be a tremendous loss of biodiversity. This excludes (excuses) us from an intervention duty.

One final question remains: Why is biodiversity important? Of course biodiversity could have instrumental value in the sense that it could contribute to the well-being of sentient beings. But if a world without predators and with a lower biodiversity would have a higher overall well-being, an instrumental value of biodiversity is not sufficient to allow predation. Hence, a mere instrumental value would violate our moral intuition in the predation case. Therefore the biodiversity solution only works if biodiversity has *intrinsic* value (here meant as being the opposite of instrumental value).

So in what sense can biodiversity (everything that is the direct product of evolutionary processes) be understood to have intrinsic value? The intrinsic value of biodiversity could be compared with the intrinsic value of well-being. A person or sentient being has two moral values: he/she is irreplaceable and he/she has a well-being which has intrinsic value. We could say that ecosystems, too, are irreplaceable and have a biodiversity that has intrinsic value.¹⁰ Let's explore this analogy between well-being and biodiversity a bit further.

1) Looking at sentient beings, we see that they tend to increase their well-being. That is because these beings have multiple needs, and they are looking for strategies to satisfy their needs as much as possible (trade-offs, resource scarcity and incompatible strategies limit their growth of well-being, though). Now, looking at ecosystems, we see that they tend to increase their biodiversity. That is because these ecosystems consist of procreating living beings, and they are subject to genetic variation (natural selection by resource scarcity limits the growth of biodiversity, though).

⁹ Universalization does not mean that only the actually living moral beings have the duty. The current number of moral beings might be too low and their skills too limited to stop all predation everywhere. Hence, intervention by the current moral agents might not endanger biodiversity. However, a duty should not depend on an arbitrary state of the world (e.g. the current state with the current number of moral agents). A duty should also be universalized in all hypothetical situations, including situations where there are enough moral agents with enough skills to prevent predation everywhere.

¹⁰ Note that – in contrast with persons – ecosystems do not have clear boundaries, so it might be problematic to speak of the irreplaceability of ecosystems. Only the whole Earth has a clear boundary as an ecosystem. Furthermore, as we have seen in section 4.6.4.3, the problem of irreplaceability of sentient beings can be solved without a need to introduce an 'irreplaceability value'.

2) Both well-being and biodiversity are a collection of different things: pleasure, friendship, absence of pain and reading a good book all contribute to well-being, just like genes, biotic landscapes, ecological processes, species and genera all contribute to biodiversity. Both well-being and biodiversity are natural properties that are difficult to calculate and express in one number, but we are able to see increases and decreases.

Note that the intrinsic value of biodiversity should be distinguished from a problematic intrinsic value of species. As species are abstract biological categories, the notion of species value faces serious problems. First, the definition of a species is complex and the moral relevance of a definition based on fertility of offspring is not clear. Second, looking at ring species and hybrids, species can have fuzzy boundaries. Giving intrinsic value to species might be as bizarre as giving value to cheeks: hitting your left cheek lowers your well-being, just as killing a species lowers biodiversity, but that does not mean that cheeks have intrinsic value.

3) It is good to increase well-being. Is it possible to increase biodiversity in nature by introducing new species through e.g. genetic engineering? The answer is no: genetic engineering contributes to the variation of life forms (as long as it does not result in increased competition and extinction of species), but it is not a *direct*¹¹ result of evolution. Genetically modified species are consciously created by intelligent beings. They are not the product of a blind process of genetic mutation and natural selection. Only the variation that is the direct consequence of natural evolution counts as biodiversity.

We can compare this with a problem in welfare ethics: the ‘experience machine’ (Nozick, 1974). Imagine a virtual reality machine that gives a lot of pleasure when your brains are plugged into the machine. Even if pleasure experiences (all positive feelings) increase, most people feel reluctant to plug into the machine, because they might have a need for authenticity (being in the real world instead of a virtual reality), or they might have a need to actually do something (getting pleasure through activity instead of through merely experiencing things).

Just as biodiversity is composed of the variation of all life forms and processes that are the direct result of natural evolution, we can say that well-being is composed of the variation of all positive feelings and emotions that are the result of preference (need) satisfaction. In this sense, the pleasures experienced in the experience machine do not contribute to well-being, unless the individual wants these pleasure experiences (i.e. if the individual has preferences for these experiences in the machine).¹²

¹¹ It is an indirect result, because the intelligence of the scientists is a result of evolution.

¹² As we have seen in section 4.2.2, this combines mental state accounts with preference satisfaction accounts of well-being (see Shaw 1999, chapter 2).

The only possible strategy to increase someone's well-being, is by eliminating obstacles that prevent preference satisfaction (i.e. eliminating barriers that enforce trade-offs, or eliminating scarcities). Similarly, the only possible strategy to increase biodiversity is by eliminating ecosystem pressures that increase competition over scarce resources. Introducing drugs (or an experience machine) to increase happy feelings is similar to introducing new species (or genetic engineering) to increase variation in life forms: they do *not* contribute to valuable well-being and valuable biodiversity.

All in all: well-being is for a sentient being what biodiversity is for a natural ecosystem. We should not lower someone's well-being without good reason, and we also should not lower biodiversity without good reason. Neither should we be willing to have a much lower biodiversity. The following table summarizes the analogy between biodiversity and well-being.

Well-being	Biodiversity
Natural property of sentient beings	Natural property of ecosystems
Tendency to increase	Tendency to increase
Constraints: trade-offs and (resource) scarcity limit growth of well-being	Natural selection: competition and (resource) scarcity limit growth of biodiversity
Variation of (positive minus negative) feelings and emotions	Variation of living organisms and processes
Result of preference satisfaction	Direct result of evolution
No 'artificial means' to increase well-being using an experience machine	No artificial means to increase biodiversity using genetic engineering

In summary, the above discussion allows us to introduce the following principle:

The intrinsic value of biodiversity. Biodiversity is composed of all variation in life forms and processes that are the direct result of natural evolution, where natural evolution is generated by random genetic mutations. We should protect biodiversity, because the biodiversity for ecosystems is analogous to well-being for sentient beings: both are intrinsically valuable properties of an entity (ecosystem, sentient being) that is unique and irreplaceable.

The above 'biodiversity principle' or 'triple-N-principle' has some resemblance with Regan's amorality criterion. But it is not about the amorality of the lion, but the

amorality of nature and evolution. This amoral nature/biodiversity criterion is a new articulation of some shared moral intuitions and attitudes of animal rights activists. They are often worried that interfering in nature is a kind of human arrogance. Perhaps this might be a bit similar to the cultural relativists who claim that imposing our human rights on other cultures is arrogant.¹³ It is not the fact that the lion is an amoral agent, which grants him the right to kill and eat others. The point is that the lion is part of a big thing, which we will call nature or the 'Other'. This 'Other', however, has a completely different morality than ours. It is 'amoral', which means: beyond the morality of our moral world. Condoning predation is a kind of respect for this Other. We are not responsible for the cruelty that evolved within the world of the Other.

If a lion decides to hunt a zebra, that is part of the amoral world. But if I decide to use sentient beings for transplantations, it is part of our moral world. As predation contributes to biodiversity whereas transplantation doesn't, this makes a distinction between predation and transplantation. Having said all this, I do believe that predation is a very serious moral problem, and perhaps some (utilitarian inclined) ethicists are right that in the end we should look for ways to intervene and decrease the vast amounts of suffering in the wild, including the suffering caused by predators. We should not be afraid of intervening, we should not be afraid of being too arrogant, because wild nature is really arrogantly cruel. At least we should not underestimate the cruelty of nature, and at least we should openly discuss this issue of decreasing wild animal suffering.

10.6 Some further tests for the 3-N principle

Let us test this biodiversity principle (or triple N principle) with some examples. Let us check whether the above hypothesis is compatible with our moral intuitions. If it is compatible with our intuitions, the animal rights ethic extended with the above principle becomes more coherent.

As a first example, suppose someone becomes ill and needs new medicines that have to be tested on sentient beings. The experimentation did not originate from an amoral process of blind evolution; it originated from moral agents. Therefore, an animal rights activist should be against experimentations on sentient beings (without their consent).

¹³ This is not an argument pro cultural relativism. Many animal rights activists are not cultural relativists.

A second example is the killing of insects by accident. Suppose that we discover that ants are sentient beings. What will happen to our ethics? Do we have a duty not to move, in order to save the ants? In other words, does the ant not only have a right not to be used as merely a means, but also a right not to be killed by us walking around? It is clear that the behavior of walking around originated by a blind evolutionary process, and that walking around is very vital for large animals like us. If we did not walk around, we would not even be here alive today. According to the hypothesis, we are then still allowed to walk around and accidentally kill ants. This conforms to our moral intuitions. On the other hand, road kill, the killing of sentient beings by cars, is morally wrong, because cars are not natural and driving a car does not contribute to biodiversity.

What about killing an annoying fly on purpose? If we would discover that flies are really sentient beings, then our hypothesis says that we should not kill the fly, because we, as aggressors, did not have a vital need that was in danger. Even if killing flies is natural and normal. And we should definitely not use flies as merely a means to our ends. So the fly would have a right not to be used as merely a means, and also the right not to be killed on purpose for non-vital needs.

Another example is the birth of animals which will have a lower value of life than ours. Imagine that all animals have a lower life expectancy and lower capacities for well-being (lower emotional richness) than humans. Those animals are like (mentally) disabled humans. From behind a veil of ignorance, you can choose between two worlds. In the first world, animals are born with a lower (but still positive) value of life. In the second world, there are only humans with a high value of life. You might prefer the second world, because in that world you have probability 1 of being a human with a high value of life. In the first world you risk being born as an animal with a lower value of life. This thought experiment would imply that it is good to make all non-human animals infertile, so that those animals go extinct.

A similar problem arises with the birth of animals that will have a high critical resource consumption level (i.e. the positive, non-zero level of resource consumption at which well-being of that individual equals zero). Total well-being would be maximized if only the beings with lowest critical resource consumption levels would procreate. The resources needed for one individual with a high critical resource consumption level can better be spent on more individuals who have the lowest critical consumption levels. The latter individuals can generate more welfare with the same amount of resources (see Shiell, 2005).

The conclusion that those animal species that do not contribute enough to the welfare function (as described in section 4.6.7.1) are no longer allowed to procreate, goes against our moral intuitions. If we say that biodiversity has moral value and that procreation is normal, natural and necessary, those animals are still allowed to procreate. As with predation, we do not have a duty to stop procreation. But we are allowed to intervene in procreation to some degree. For example if parents know that

their future child will be disabled (if the potential child would not contribute to the welfare function), then (early) abortion would be allowed.

Related to the above problem is the issue of genetic enhancement. Do we have a duty to genetically enhance species who would otherwise not contribute enough to the welfare function? Should we genetically modify frogs to increase their potential levels of lifetime well-being? Or genetically modify lions such that they no longer need meat? Changing genes on purpose is not a natural behavior. Hence, those new genes do not contribute to biodiversity. If we would replace all frogs with enhanced frogs, some biodiversity will get lost because the genes that characterize the unenhanced frogs will disappear from the gene pool and the new genes do not contribute to biodiversity. The question is whether this loss of biodiversity (the loss of some specific genes) trumps the increase of the welfare function. This becomes an intuitive balancing of two competing values which can be approached in a democratic way (see section 4.6.11).

A final example is the situation where a child holding a gun is about to kill another child. This child has no moral agency, but still we have the duty to intervene and protect the second child. That is because the first child does not have a vital need to kill the other child. Also, killing with a gun did not evolve by an evolutionary process, so it is not natural.

The above examples indicate that our triple-N-principle is coherent with a lot of our moral intuitions. As a summary, the following table gives an overview of solutions for five of the abovementioned challenges to a consequentialist welfare ethic.

1. Predation: carnivores are allowed¹⁴ to hunt, kill and eat many prey animals, even if they harm sentient prey.
2. Motion: a human (or another big animal) is allowed to move around and kill (by accident) many insects, even if insects were sentient.
3. Organ transplantation: a surgeon is not allowed to sacrifice a victim without informed consent, even if patients are dying when they do not receive new organs.
4. Medical experimentation: a researcher is not allowed to sacrifice someone without informed consent, even if the developed medicines could save many patients in the future.
5. Procreation: all animals are allowed to procreate¹⁵, even if those animal species do not contribute enough to welfare function.

¹⁴ In the sense that we (moral agents) do not have a duty to interfere in predation to save the prey.

¹⁵ This does not imply that they are allowed to have as many offspring as they want, because we have to avoid overpopulation. For example having more than 3 children would be problematic if this rule would be universalized: the resulting exponential population growth will hit the boundaries of the planet.

	Intuitive moral judgments	Consequentialist welfare ethics	Mere means principle	Principle X
1. Predation	Allowed	Not allowed	Not allowed	Allowed
2. Motion	Allowed	Not allowed	- (undecided)	Allowed
3. Organ transplantation	Not allowed	Required	Not allowed	- (undecided)
4. Medical Experimentation	Not allowed	Required	Not allowed	- (undecided)
5. Procreation	Allowed	Not allowed	- (undecided)	Allowed

A common property of the above problems is the *necessity of a serious welfare loss*. The behavior (predation, transplantation,...) is necessary for:

1. existing beings to stay alive (the predators, patients and big animals in problems 1 to 4),
2. potential beings to have a life¹⁶ (problem 5), or
3. populations to survive (problem 5).

The second column in the above table presents the intuitive moral judgments about the allowance of the five types of behavior. This column is the opposite of the third column, which gives the results according to consequentialist welfare ethics. To reconcile the welfare ethics with the moral intuitions, we can develop a two-step approach.

The first step (column four) introduces the deontological Mere Means Principle (discussed in section 6.2), which changes a few consequentialist judgments (organ transplantation and medical experimentation). This principle adds some impermissibilities that trump the consequentialist principle, so after this step, none of the five behaviors is allowed. But that is not sufficient to match all judgments with the second column in the table.

Therefore, in a second step, a new principle X has to be introduced (column five). This principle X trumps both the mere means and the consequentialist principles. After this second step, the (im)permissibility of the five behaviors matches the intuitions. We have seen that the 3-N-principle, which corresponds with the value of biodiversity, serves as a good principle X, because it changes the judgments in the three remaining problems: predation, motion and procreation. There are different ways to measure biodiversity. Scientists can propose different biodiversity metrics. But each metric shows a drastic decrease when predation, motion and procreation stops. Hence,

¹⁶ Procreation is necessary for a potential being to get a life, in the sense that without procreation, the potential being could never come into existence.

whatever biodiversity metric one chooses, it is compatible with the moral intuitions of a deontological naturalistic ethic that permits those types of behavior.

The intrinsic value and the resulting 3-N-principle fit in a wide reflective equilibrium (Daniels, 1979): they are coherent with strong moral intuitions in three different cases (predation, motion and procreation), they are coherent with a notion of naturalness (which a lot of people care about), and they are coherent with some background theories of biodiversity (some properties of biodiversity make the above mentioned analogy between biodiversity and well-being sensible, the latter having intrinsic value in consequentialist welfare ethics).

The next section introduces a second possible principle X. Also this principle fits in a wide reflective equilibrium: it is coherent with the same three moral intuitions, and coherent with a notion of fairness.

10.7 A second hypothetical solution to the difference problem: behavioral fairness

The above 3-N-principle refers to some seemingly arbitrary criteria of naturalness and normality, and to some seemingly mysterious intrinsic value of biodiversity (it is an intrinsic value that a moral agent gives to a property of a non-sentient entity). For those who feel uncomfortable with the above solution, there is a second promising solution to the difference problem that avoids these references to nature and biodiversity. This second solution is based on a notion of fairness. If the lion is not allowed to eat the zebra, then the lion could say that as a matter of fairness, the zebra is not allowed to eat either. So the principle claims that A is allowed to do X with B, if B also does X and is allowed to do X. If the zebra (B) is allowed to eat (do X), then the lion (A) is allowed to eat as well. So if the zebra eats something, and if the zebra would say that she is allowed to eat, then the lion is allowed to eat as well, even when the zebra is the food. Similarly: if a (sentient) insect is moving around and is allowed to move around, then I am allowed to move around as well, even if that harms the insect by accident.

This is in contrast with organ transplantations and medical experiments. I cannot use you as merely means for organ transplantation or medical experiments, because you don't perform medical experiments and transplantations yourself. So my claim that I am allowed to use you becomes invalid, because you are innocent when it comes to medical experiments. The zebra and the insect, on the other hand, were not innocent when it comes to respectively eating and moving.

The tricky point is: what is the behavior X exactly? The zebra could say to the lion: “You are not allowed to eat me.” Then X could refer to ‘eat this zebra’. Fairness requires that the lion could say to the zebra: “Then you are not allowed to eat yourself either.” But that’s fine for the zebra: she was already following the rule not to eat herself. Similarly, if X meant ‘eat zebra’, the zebra could live with the rule that it is not allowed to eat zebras. The same goes for X equal to ‘eat animals’. The idea is that the behavior X should not refer to specific individuals or groups of individuals. X should only refer to a behavior, such as ‘eat’ or ‘do medical experiments’. If the lion is not allowed to eat, then neither is the zebra, and the zebra could not live with that rule.

But then another problem lurks. To what kind of behavior should X refer? You could say: “You are not allowed to use me in experiments, because I’m innocent: I don’t use anyone in experiments.” My reply could be: “Sure, but you do use plants for food, so the true X means ‘use someone or something’”, and according to that view, you are guilty. But that X would be too general: it would imply that all kinds of uses are not allowed. Therefore, X should be the most accurate and specific description of a behavior.

The most accurate description of the behavior of the lion would be: ‘eat’. But if you are allowed to eat (e.g. eat plants), then I would be allowed to eat you. In particular, sometimes you do eat plants just for taste. If a being is allowed to eat something for taste (instead of survival), then so do I, and if that being happens to taste good, I am allowed to eat her? I am allowed to eat you if you taste good and if you eat something for taste?

To avoid this conclusion, we have to refer to the intention or the purpose of the behavior. In particular, we can keep the necessity criterion of the above 3-N-principle. Therefore, X should be more specific, such as ‘eat for survival’. I can survive without eating you, so eating you would not be for survival.

It is not always clear to see whether someone eats for survival or pleasure, but if someone eats a sentient being, causing harm to the sentient being, and if there are healthy alternatives available, the intention to eat for survival will be flawed.

So far, we have a promising fairness principle which roughly sounds as: “If you are allowed to do a specific behavior for survival, then so am I allowed to do the same type of behavior for survival.” But this principle still needs some further refinements and clarifications when we look at more hypothetical cases.

First: what if plants were sentient beings? Plants don’t kill and eat other living beings. Does this mean that plants are innocent and that herbivores (including we) should not eat them? Of course, plants do consume chemical resources and energy. So we have to state that “consuming chemical resources and energy” is the same type of behavior as “eating”.

Second, what if a sentient being consciously decided: “In consuming chemical resources, I will only consume non-sentient beings.”? Is it fair to eat this sentient being? This sentient being is in a sense lucky that she can survive on consuming non-sentient

beings. A lion, on the other hand, is not that lucky: he needs sentient beings. So the lion is allowed to eat this sentient being. However, if the lion was lucky in the same way (if he could survive by eating non-sentient beings), then he should restrict his consumption to non-sentient beings as well.

A third objection is based on the common-sense judgment that doing experiments on (non-sentient) plants is allowed. But now a mad scientist could say: "If you are allowed to do experiments, on plants, I am allowed to do experiments as well, on you!" To avoid this, we make a distinction between moral and amoral beings. The fairness claim cannot be put forward against the victim, when the victim is a moral being. Note that amoral humans (babies and mentally handicapped humans) and non-human animals do not perform experiments on plants, so they should not be used by the mad scientist either.

In summary, we get the following principle.

The principle of behavioral fairness. An agent is allowed to do a specific type of behavior that causes harm to victims¹⁷, if (1) the behavior is necessary, (2) the harm is minimal (the agent does not have an alternative that causes less harm), (3) the victims are amoral agents, (4) the victims perform the same type of behavior and (5) the victims are allowed to do that behavior. If the agent has a better option (that causes less harm), then s/he should choose that option.

This fairness principle means that if A is allowed to do something, then so is B, under certain conditions. The reader can verify that this fairness principle is able to withstand the tests mentioned in the previous section: impermissibly using sentient beings against their will in experiments, permissibly getting offspring who insufficiently contribute to the welfare function¹⁸, permissibly killing insects by accident when making a movement, impermissibly killing an annoying fly on purpose, and permissibly eating plants even if plants are sentient.

¹⁷ Here, causing harm should be understood in a broad sense. E.g. lowering the total welfare function is a cause of non-personal harm where the victims can be considered as the total population. Also using someone as merely a means is a cause of harm.

¹⁸ Procreation is not a necessary need for an existing individual: no-one will die without procreation. But it can be considered necessary for potential future beings.

10.8 Summary

We have seen three principles, the fairness principle, the triple-N-principle (or biodiversity principle) and the tolerated choice equality that solve the prey problem and the difference problem, the two components of the predation problem. Those principles can help us derive a consistent and coherent animal rights ethics that can be reconciled with moral intuitions shared by a lot of people.

Regarding the 3-N-principle, I first clarified the meaning of normal, natural and necessity. Second, I demonstrated that none of the three criteria separately are sufficient. Third, I showed that the combination of all three criteria can make a distinction between predation and transplantation. Next, I made a connection between the three N-criteria (as we have defined them) and biodiversity. Fifth, I explained in what sense biodiversity can be said to have moral value. And finally I tested the 3-N-principle in other situations, indicating that it corresponds with our moral intuitions. The result is a principle that says that basic right violations are only allowed when all three N-criteria are met.

The three N justifications (normality, naturalness and necessity) that animal rights activists seem to hold are specific interpretations of the same three justifications in the ideology of carnism (Joy, 2009) and might give us a further clues about the question why it is so difficult to convince people to become vegetarians or vegans. Becoming vegan should be easy in principle, because eating animal products is not necessary for survival. Yet, carnist people appear to have strong emotional objections, and they use a lot of naturalistic fallacies to justify their behavior. For example, in the US there are more people in favor of hunting for pleasure, than there are in favor of medical experiments with animals (Herzog, 2010). And in contrast to hunting, medical experiments are believed to save lives. It becomes clear that there are some hidden sensitivities that people – including animal activists – might have. We mentioned the fear for human arrogance and the apparent similarities with cultural relativism and respect for the Other. But a belief in naturalness might be an important moral intuition of a lot of people.

The 3-N-principle can be justified with a reference to an intrinsic value of biodiversity. However, those who dislike such a mysterious intrinsic value could rather adopt another principle to solve the predation problem: the behavioral fairness principle. This principle says that a specific behavior that harms a victim (e.g. killing and eating a living being) is allowed only if the action is necessary for survival, if the victim (e.g. the prey) is an amoral being, if the victim is also guilty of a similar type of action and if s/he is allowed to do that action. If the zebra eats for survival and is allowed to do so, then so also is the lion. But if a harm (a loss of well-being) occurs to a

sentient victim, and if there are healthy alternatives that do not involve such harm, then of course those alternatives should be chosen.

Finally, we can combine the 3-N-principle and the principle of behavioral fairness into a new equality principle: everyone has an equal right to a behavior that is both natural, normal and necessary (i.e. a behavior that strongly contributes to biodiversity).

Chapter 11 The property problem and the harvest problem

As animal products are not necessary for us, we (humans) should become vegans. But veganism is not good enough. There is another issue: what about property rights for animals? In particular: are we allowed chasing away animals to clear a forest in order to build a house or extend a crop field? What about the birds and the squirrels who built nests in those trees? We destroy their houses and habitats by cutting down the trees. From an antispeciesist point of view, this is similar to the destruction of someone's property or stealing someone's land.

And what if we use a crop field (to produce vegan food)? The crop field is our property, but some other animals invade the field and start eating our grains, fruits and vegetables. Are we allowed to chase them away, even if this results in more suffering by those animals due to an increased competition between those animals for scarce food resources? Are we allowed to kill them if they keep returning to our crop fields?

And third, what if we kill those small animals by accident when we use machines to harvest our crops or drive with big trucks? Some animals (small rodents, birds,...) die, even when vegan food is produced. At www.animalvisuals.org, an estimate is given of the number of animals killed due to harvesting. For a vegan it comes down to roughly 2 animals killed per year. This is still much less than the number of animals killed on purpose and by accident in the production of animal products (meat requires slaughtering an animal but also harvesting crops for livestock feed), but it is not something we can dismiss so easily.

11.1 Habitat destruction

Let us consider the first problem: the destruction of wild animal habitats. Indeed, we are not allowed to invade a foreign country, destroy the houses and chase the local people away. Isn't it speciesist to allow the destruction of animal habitats, destroying the nests of the birds and squirrels? Or do animals have habitat rights?

A consistent ethical theory of animal equality indeed implies that animals have habitat rights similar to our property rights. But this does not mean that we are not allowed to use natural resources even when using those resources requires invading someone's habitat. We are allowed to use habitat of animals to some degree, because if we are not allowed to use natural resources, then neither is no-one allowed. If an animal uses natural resources and a habitat, it means that another animal is no longer able to use those resources. But animals are allowed to use natural resources, and hence so are humans. As habitat and natural resources are scarce, there is always competition between different sentient beings.

To demonstrate that using an animal's habitat is not necessarily speciesist, imagine that a mentally disabled human escaped from a care institution. This human is as intelligent as a bird, and he decides to climb into a tree and build a nest. As a tree is a bird's habitat, this tree becomes the habitat of the disabled human. Unfortunately, we want to cut down that tree, to produce paper for important books, or to extend our cropland, or to build a house for ourselves. Chasing away the disabled human can be considered as a harm to that human. The mentally disabled human might get injured, or he might run to another tree that is already occupied by another sentient being (say another escaped mentally disabled human), increasing their competition for scarce trees. This harm done to the mentally disabled human is not a use as merely a means. Chasing away the human does not violate his basic right.

We are allowed to cause harm to someone, as long as the victim is not used as merely a means, and as long as the quasi-maximin prioritarian principle of justice is not violated. In practice, this latter condition means that 1) we should be very careful in cutting down trees and harvesting crops, 2) we should strongly decrease our consumption of natural resources in order not to invade too much in someone else's habitat, 3) we should stop the further destruction of wild animal habitat (increase the area of nature reserves and stop the expansion of human settlements) and 4) we have a strong responsibility to help potential victims (taking care of wild animals by strongly increasing our support for wildlife rescue centers, giving food aid to animals in need, protecting wild ecosystems). Donaldson and Kymlicka (2011) developed a political theory of animal rights, whereby the habitat of wild animals should be considered as sovereign nations or sovereign animal territory. We do not have the right to colonize and displace the citizens (wild animals) of these spaces. If we move into animal

territory, we should compensate any harm done to wild animals, by helping wild animals in need. Risks to wild animals can be compensated with benefits for them. Therefore, merely veganism (abstaining from consuming animal products) is not good enough. A vegan does not violate the basic right not to be used as merely a means, but s/he might still violate the quasi-maximin principle if s/he is not careful. It can be compared with invading another country: even if you are not killing humans for food (you are anticannibalistic), you do cause harm when you steal someone's land.

If we are allowed to carefully chase away wildlife animals to some degree, and if we want to avoid speciesism, we are also allowed to chase away those escaped tree sitting mentally disabled humans. In a sense, this means that mentally disabled humans and wildlife animals have weaker property rights than rational beings like you and me (i.e. mentally capable humans). It can be argued that this difference in the strength of property rights can be derived from the quasi-maximin principle, if we keep in mind that differences in mental capacities generate differences in how property rights influence well-being. In other words: some sentient beings have special mental capacities, which means that their well-being is more strongly influenced by how properties are distributed. This happens especially when a sentient being has the capacity to understand the notion of a property, to invest in his/her property (making him/her feel more emotionally attached to his/her property) and is able to cooperate with others in the search for a fair distribution of property. Those people (e.g. rational beings) might have stronger property rights compared to other sentient beings, because their well-being is more strongly dependent on the distribution and treatment of properties.

Therefore, there is a difference between destroying the house of a rational being and destroying the nest of a bird by accident in the case of cutting down a tree. First, a more rational being might have a stronger emotional connection and understanding of property, and second, a house is different from a nest in terms of effort to construct it and in terms of replaceability. Therefore, it may be worse to violate (mentally healthy) human property rights than (non-human) animal habitat rights.

This stronger property right for rational beings can also justify why rational beings are more strongly permitted to defend their own properties (e.g. defend their crop fields against invaders). Yet, we should look for animal friendly, non-lethal methods to avoid animals eating our crops, just as we have to look for human friendly solutions when a group of humans invade our cropland to steal our food. Killing them or poisoning them can only be a very last resort. And some solidarity with animals might also be required to compensate for some harm done to wild animals: we should be willing to produce food to help hungry (wild) animals. Again, this means we should more strongly support wildlife rescue centers.

In this context, we can also refer to a principle of tolerated partiality. We might feel more concern to respect the property rights of mentally disabled humans compared to

the property rights of birds, but we should tolerate someone who takes more effort to respect bird habitat above the property of a mentally disabled human.

11.2 Animals killed in harvest

Imagine there are two animals (e.g. small rodents) that are killed in agriculture to produce vegan food for one human for one year. Are we allowed to farm some land if we know that per vegan person every year on average two vertebrate animals are killed? Or is vegan agriculture a violation of the quasi-maximin prioritarian principle?

To shed some light on this issue, remember first of all that in section 4.2 we saw two factors that influence someone's lifetime well-being: 1) the richness of emotions (some sentient beings have more and stronger preferences and can experience higher levels of momentaneous well-being if those preferences are satisfied) and 2) psychological connectedness with someone's past and future (some sentient beings have richer and stronger autobiographical selves as well as stronger preference towards the future).

Imagine we have four beings: a human person (with high mental capacities and a strong psychological connectedness), a big animal (e.g. a cow) and two small animals (e.g. rodents). There are three situations: 1) the human produces vegan food, at the cost of endangering the two rodents (those two rodents might get killed if they happen to be in the crop field at the wrong time), 2) the human does not produce vegan food nor kills the cow, which means the human cannot eat and the rodents and cow survive, and 3) the human eats the cow, the rodents are not in danger because the cropland is not harvested. Which of these situations is the best from the point of view of prioritarian justice?

As the rodents likely have lower potential levels of momentaneous well-being and also have a weaker psychological connectedness with their past and future selves, they likely have lower levels of integrated lifetime well-being. Looking at the welfare function equation in section 4.6.8, it is far from obvious that situation 2 is better than situation 1. After all, the lifetime well-being of the vegan is strongly reduced when s/he is no longer allowed to farm some land, and the early death of the rodents does not strongly decrease their integrated lifetime well-being if those rodents have less psychological connectedness.

What about situation 3? This situation is definitely the worst with respect to the basic right principle (the basic right of the cow is violated in situation 3), but is it also the worst with respect to lifetime well-being? The welfare function does not give a clear answer, except that one might think that the death of one animal (the cow in situation 3) is better than the death of two animals (the rodents in situation 1).

The principle of prioritarian justice says that we should not cause more harm in such a way that the welfare function decreases. But it is far from obvious which of the above three choices is the best from the point of view of prioritarian justice. In that case, we can do two things.

First, we can refer to the principle of tolerated partiality. That principle was first derived in situations where we help others, and the choice is between helping A versus helping B. Now we are facing a situation where we are harming others, and the choice is between harming A versus harming B. The principle of prioritarianism says that we should minimize harm, where harm is now defined as a decrease of the welfare function. But if it becomes difficult to calculate whether the welfare function decreases, we can say that we are allowed to be partial to some degree, to prefer the choice that is in our own best interest, i.e. situation 1.

Second, we can introduce a heuristic rule of thumb: harming an identifiable victim is worse than harming a non-identifiable victim. The animals that die by accident in agriculture (e.g. the two rodents) are non-identifiable victims. If I eat a vegan product, I cannot identify the rodents that died due to harvest. It might be possible that no animal was harmed when a vegan meal was produced. On the other hand, if I consume an animal product (e.g. the meat of the cow), I do know that this product comes from an identifiable victim. A piece of meat, an egg or a drop of milk comes from someone's body, so at least someone is harmed, and we know who. In that sense, situation 3 is worse than situations 1, because the cow is an identifiable victim that is killed whereas in situation 1 there is no identifiable harm (it is not clear that the two rodents are actually killed). Also in situation 2, there is an identifiable harm to the vegan human who has nothing to eat. Therefore, situations 2 and 3 are worse than situation 1, according to this heuristic rule of identifiable harm, because situations 2 and 3 involve identifiable victims.

Furthermore, we can add that the harm done to non-identifiable victims can be compensated by helping animals who need help. If my behavior results in the death of one non-identifiable animal (i.e. I don't know who dies, I only know that someone dies), I can compensate this harm by saving the life of an identifiable animal (i.e. a specific animal who needs help). Even if harming identifiable victims is not worse than harming a non-identifiable victims, there is still a difference: the second harm can be compensated by helping others who we did not harm, whereas the first harm can only be compensated by helping the harmed identifiable victims themselves. As using resources (e.g. doing agriculture) and emitting waste harms non-identifiable animals, we can compensate for this harm by sufficiently helping other animals.

In summary: vegan agriculture might be permissible, even if non-human animals die by harvesting the crops. What it means is that we should first of all be much more careful in agriculture, mining and forestry, trying to avoid harming and killing animals (e.g. permaculture and zero tilling agriculture). Second, we should lower our ecological

footprint and lower our use of natural resources such as cropland. And third, we should compensate harm done to wild animals by strongly supporting wildlife rescue centers that help all kinds of wild animals (including small rodents and birds). Merely eating vegan food is not good enough according to a consistent ethic of animal equality.

Part 4 Epilogue

Chapter 12 The moral hands

12.1 The moral hand of normative ethics: five principles of a complete and coherent ethic

The previous chapters discussed a lot of ethical principles. As a summary, and to demonstrate the compactness, completeness and coherence of the ethic, let me use a metaphor: the moral hand. Each of the five fingers represents an ethical principle.

-**The thumb:** rule universalism (see section 1.2). You must follow the rules that everyone (who is capable, rational and informed) must follow in all morally similar situations. You may only follow the rules that everyone (who is capable, rational and informed) may follow in all morally similar situations. Prejudicial discrimination is immoral. We should give the good example, even if others don't. Just like we have to place the thumb against the other fingers in order to grasp an object, we have to apply the principle of universalism to the other four basic principles.

-**The forefinger:** justice and the value of lifetime well-being (see section 4.4). Increase the well-being (over a complete life) of all sentient beings alive in the present and the future, whereby improvements for the worst-off positions (the worst sufferers, the beings who have the worst lives) have a strong priority. Lifetime well-being is the value you would ascribe when you would live the complete life of a sentient being, and is a function of all positive and negative feelings that are the result of (dis)satisfaction of preferences (everything wanted by the being).

-**The middle finger:** the mere means principle and the basic right to bodily autonomy (see section 6.2). Never use (or consider) the body of a sentient being as merely a means to someone else's ends, because that violates the right to bodily autonomy. The two words 'mere means' refer to two conditions, respectively: 1) if in order to reach an end (e.g. saving someone) you force a sentient being to do or undergo something that the being does not want, and 2) if the body of that sentient being is necessary as a means for that end, then you are not allowed to treat that being in that way. A sentient being is a being who has developed the capacity to want something by having positive and

negative feelings, and who has not yet permanently lost this capacity. They have the experience of having preferences. The middle finger is a bit longer than the forefinger, and so the basic right is a bit stronger than the lifetime well-being (e.g. the right to live). The basic right can only be violated when the forefinger principle of well-being is seriously threatened.

-**The ring finger:** naturalness and the value of biodiversity (see section 10.4). If a behavior violates the forefinger or middle finger principles, the behavior is still allowed (but not obligatory) only if that behavior is both natural (a direct consequence of spontaneous evolution), normal (frequent) and necessary (important for the survival of sentient beings). Just as lifetime well-being is the value of a sentient being, biodiversity is the value of an ecosystem and is a function of the variation of life forms and processes that are a direct consequence of natural evolution. The valuable biodiversity would drastically decrease if a behavior that is natural, normal and necessary was universally prohibited (universally, because you have to put the thumb against the ring finger).

-**The little finger:** tolerated partiality and the value of personal relationships (see section 5.1). Just as the little finger can deviate a little bit from the other fingers, a small level of partiality is allowed. When helping others, you are allowed to be a bit partial in favor of your loved ones, as long as you are prepared to tolerate similar levels of partiality of everyone else (everyone, because you have to put the thumb against the little finger).

These five fingers might look like post hoc rationalizations of moral intuitions (see e.g. Haidt, 2001; Greene, 2008), and in fact they are. Except that each of the five principles is backed up by a coherent set of arguments, which makes it less of an arbitrary system.

The thumb principle allows us to satisfy a need for consistency and impartiality (non-arbitrariness), it allows us to deal with game-theoretic ideal and non-ideal situations (with or without universal compliance, see section 7.2), it allows us to tackle moral illusions, it fits with our intuition that one should give the good example, and it helps to generate principles (other principles such as the ring finger and little finger only work when the thumb principle is applied).

The forefinger can be derived by either a veil of ignorance thought experiment (assuming some levels of risk aversion, loss aversion and uncertainty aversion, as well as some aspects of prospect theory, see section 4.6) or by the moral virtue of compassion (assuming a need for a certain level of well-being efficiency, as well as some moral intuitions about variable populations and personal identity). The lifetime perspective is justified by two coherent reasons: the intuition that persons are not replaceable, and the intuition that there is a difference between intrapersonal and interpersonal harm, i.e. a difference between imprudent behavior (harming your future self) and immoral behavior (harming another person).

The middle finger is consistent with moral intuitions in many (at least ten) dilemmas, with notions of respect and intrinsic value, and with a propertarian libertarian ethic of bodily autonomy.

The ring finger fits with moral intuitions in situations of predation, procreation and motion, and is compatible with moral intuitions about naturalness (as used in e.g. a carnist ideology) and the value of biodiversity in environmental ethics.

The little finger is compatible with our intuitions about partiality in situations of imperfect, positive and procreational duties, and it fits in an extended mere means principle of the middle finger (section 6.6).

Furthermore, the principles are also made compatible with scientific background theories of biology (e.g. gradual evolution) and psychology (e.g. personal identity and mental capacities). These background theories challenge some common sense assumptions on e.g. boundaries between species, boundaries between persons and boundaries between levels of mental capacities. Our moral hand is perfectly able to deal with non-sharp boundaries between species, persons and mental capacities. For example the mathematical expression of the forefinger principle includes a connectivity function to deal with complex issues of psychological connectedness and personal identities (section 4.6.8), the basic right of the middle finger couples gradations of mental capacities with gradations of needs (section 6.4), and the biodiversity principle of the ring finger works well even when there are no clear boundaries between species.

These considerations challenge the objection that the ethic of the five fingers is merely a post hoc rationalization of intuitions. It may be a rationalization, but it is also a coherent system in wide reflective equilibrium (Daniels, 1979). It is a mystery why a rationalization of moral intuitions and emotions (Greene, 2008) can have such a level of coherence.

12.1.1 Five principles of equality

The five fingers of the moral hand of normative ethics produce five principles of equality.

-**The thumb:** the formal principle of impartiality and antidiscrimination. We should treat all equals equally in all equal situations. We should not look at arbitrary characteristics linked to individuals. This is a formal principle, because it does not say how we should treat someone.

The other four principles are material principles of equality. They have specific content and are generated when the thumb is applied to the four fingers.

-**The forefinger:** prioritarian equality of lifetime well-being (the principle of priority for the worst-off). As a result of this priority, we have an egalitarian principle: if total

lifetime well-being is constant between different situations, then the situation which has the most equal distribution of well-being is the best.

-**The middle finger:** basic right equality. All sentient beings (with equal levels of morally relevant mental capacities for well-being) get an equal claim to the basic right not to be used as merely a means to someone else's ends.

-**The ring finger:** naturalistic behavioral fairness. All natural beings (who contribute equally to biodiversity) have an equal right to a behavior that is both natural, normal and necessary (i.e. a behavior that contributes to biodiversity). Natural beings are beings evolved by evolution. E.g. if a prey is allowed to eat in order to survive, a predator is allowed to do so as well (even if it means eating the prey). If the natural, normal and necessary behavior involves several options, the option that causes the least harm (the least loss of well-being, the least violations of basic rights and the least loss of biodiversity) should be chosen (e.g. if an omnivore can survive by eating sentient animals as well as by eating non-sentient beings, s/he should not eat the sentient beings).

-**The little finger:** tolerated choice equality. Everyone is allowed to be partial to an equal degree that we can tolerate. If you choose to help individual X instead of individual Y, and if you tolerate that someone else would choose to help Y instead of X, then X and Y have a tolerated choice equality (even if X is emotionally more important for you than Y).

The forefinger, middle finger, ring finger and little finger correspond with resp. a welfare ethic, a rights ethic, an environmental ethic and an ethic of care.

12.1.2 Applications of the five fingers

12.1.2.1 The fingers applied to the consumption of animal products

The five moral fingers can be applied to the production and consumption of animal products (meat, fish, eggs, dairy, leather, fur,...).

-**The forefinger:** compared to humans, livestock animals are in the worst-off position due to suffering and early death. The loss of lifetime well-being of the livestock animals is worse than the loss of well-being that humans would experience when they are no longer allowed to consume animal products. Livestock and fisheries violate the forefinger principle of well-being.

-**The middle finger:** the consumption of animal products almost always involves the use of animals as merely a means, hence violating the mere means principle of the middle finger.

-**The ring finger:** animal products are not necessary for humans, because well-planned vegan diets are not unhealthy (according to the Academy of Nutrition & Dietetics, ADA, 2009). Biodiversity will not decrease when we would stop consuming

animal products (on the contrary, according to UN FAO the livestock sector is likely the most important cause of biodiversity loss). Hence, the value of biodiversity cannot be invoked to justify the consumption of animal products.

-**The little finger:** we would never tolerate the degree of partiality that is required to justify livestock farming and fishing. Hence, tolerated partiality cannot be invoked to justify the consumption of animal products.

It follows that veganism is ethically consistent, and the production and consumption of animal products are ethically inconsistent.

-**The thumb:** give the good example, even when other people continue consuming animal products. From this principle, it follows that veganism is a moral duty.

12.1.2.2 The fingers applied to the problem of abortion

The above section nicely demonstrated that all the five fingers are relevant in the problem of consumption of animal products. Most other ethical issues only require one or two of the five principles (e.g. only the thumb and forefinger are sufficient to argue for gay marriage). But there is another ethical issue that can only be grasped with all five fingers: the problem of abortion. This problem nicely illustrates how the five fingers work, so as an illustration let us apply the different fingers to the problem of abortion.

-**The forefinger:** early abortion is allowed, when the fetus is not yet sentient. Once a fetus has developed the capacity to feel, it becomes a sentient being. Aborting this sentient being will result in a very low lifetime well-being (value of life) due to the short lifespan of the aborted fetus. The fetus is in the worst-off position, so should get a strong priority for an increase of lifetime well-being. That means that late abortion (when the fetus is sentient) is not allowed.¹

¹ There are exceptional cases, e.g. when the future child will be seriously disabled, or when the pregnancy involves serious health risks. These have to be taken into account in the prioritarian weighing of well-being. Also, a fetus has a very low psychological connectedness with his/her future (see section 4.2.5). Some impartial observers behind a veil of ignorance might therefore have different estimates of his/her lifetime well-being (value of life) or ascribe a low connectivity function between the fetus and his/her future momentaneous minds. This means that extending the life of the fetus should not get such a strong priority. (See also McMahan, 2002.) Yet, abortion not only shortens the life of the fetus, but also prevents the existence of all future momentaneous minds and persons that the fetus could become. When these possible future persons can no longer contribute to the welfare function, the welfare function might decrease too strong. More mathematically: when a fetus is aborted at time t , its integrated well-being $\hat{\mu}_{\pi(t)}$ is a little bit lower than without abortion (it is only a little bit, due to low connectivity). As a result, this decrease of integrated well-being due to abortion decreases the welfare function a little. But with abortion, something else happens that more strongly decreases the welfare function. Without abortion, the fetus will become future persons, and hence the welfare function also includes an integral over all those future integrated levels of well-being $\hat{\mu}_{\pi(t')}$ for $t' > t$. When abortion is performed, those future integrated levels of well-being no longer count, and this

-**The middle finger:** if a woman does not want the pregnancy, and as her body is necessary for the fetus to survive (and if the woman is not responsible for this dependency, e.g. when she was raped), we can say that the fetus (unconsciously) uses the pregnant mother as merely a means. If the woman aborts her fetus, we cannot condemn her without considering her as merely a means for the ends of the fetus. This violates the extended mere means principle. So we should at least tolerate abortion, even if the fetus is already sentient (and even if the fetus would have higher mental capacities such as rationality). This line of reasoning is also reflected in Thomson's argument of the famous violinist to defend abortion (Thomson, 1971). Imagine that you are kidnapped by music lovers who connected your body to an unconscious very famous violinist. The violinist has a special disease that takes nine months to cure, and you are the only person who can save him. If you unplug yourself from the violinist, he will die. Most people have the intuition that unplugging is permissible.

-**The ring finger:** procreation is natural. One reply to Thomson's defense of abortion (the example of the famous violinist) is based on the natural-artificial distinction. According to Parks (2006), pregnancy and procreation are natural, and this is different from the artificial treatment to save the famous violinist. As procreation is natural, normal and necessary, one might say that the fetus is allowed to use the mother as merely a means.

-**The little finger:** some level of partiality is allowed. As prey are allowed to defend themselves against being used by predators (even if that implies the death of the predator), we can state that also pregnant women are allowed to defend themselves against being used by fetuses. A doctor is allowed to be partial and to choose for the woman (i.e. help to perform abortion) as well as for the fetus (i.e. refuse abortion).

The result of balancing the above four principles (the four fingers) is that early abortions are permissible (and the doctor is allowed to refuse to help). When the pregnancy is already very advanced, abortion might not be permitted. If e.g. the pregnancy takes only a few more days, the extra use as merely a means during those few days will be low. It will be the case that the violation of the mere means principle (the middle finger) becomes so low, that it will be canceled by the naturalness principle (the ring finger) which permits the use as merely a means for the ends of the fetus. After this cancelation, what is left are the forefinger and little finger principles. The priority for the lifetime well-being of the fetus (the forefinger) will be very high, so it might become

results in a much stronger decrease of the welfare function. In other words: an abortion does not strongly harm the fetus (lowering its integrated well-being $\hat{\mu}_{\pi(t)}$ a little bit), but it can still strongly harm (decrease) the welfare function as a whole. This strong decrease of the welfare function counts as an impersonal harm (McMahan, 2009).

too partial to choose for abortion (the little finger principle might be too weak to justify abortion).

12.1.2.3 The fingers applied to the environmental problem

As a third example, we can look at what the moral hand says about the twin environmental problem of overconsumption and overpopulation. Imagine there is a planet, the Earth, that contains both moral agents and amoral sentient beings. The moral agents are the people who can reflect on their own behavior and can have a strong influence on their own consumption and reproduction levels. On this planet Earth, the moral agents appear to have very high potential levels of lifetime well-being, because they have a rich emotional life, a long lifespan and a high psychological connectedness. Those moral agents also typically use a lot of resources (not only the use of material resources, but also the use of the ecosystem's absorption and processing capacities for emitted substances).

We can write an equation for the environmental impact generated by a group of moral agents. The impact (I_m) is the product of four factors: $I_m = P \cdot A \cdot C \cdot T$, where P equals the population size (the number of moral agents in the population), A equals the average affluence (or average lifetime well-being of those moral agents), C equals the average consumption level of resources per unit of affluence and T is a technology factor which equals the average environmental impact per unit of resource consumption (this factor is determined by the choice of technology).²

Ecosystems are very complex, but scientists have derived some rules of thumb to determine the effects of a high environmental impact generated by resource consumption. Some useful rules of thumb are the footprint indicators. For example the ecological footprint measures the use of bioproductive area (GFN, 2010), and the carbon footprint measures the emissions of greenhouse gases. Those footprint indicators each have a corresponding Earth's carrying capacity E . These carrying capacities, for example the total available bioproductive area or the Earth's capacity to absorb greenhouse gases, are finite. If the footprint is higher than the corresponding carrying capacity (i.e. if $I_m > E$), then we can expect that the current population of moral agents has a negative influence on the lifetime well-being of future populations of moral agents, as well as on current and future living amoral sentient beings.

-**The forefinger:** a lot of current moral agents should decrease their consumption and reproduction levels. Looking at the welfare function, we see that there is a current living population of moral agents who have a high level of lifetime well-being, but this population also generates an environmental impact that is higher than the carrying

² This equation is inspired by the famous IPAT equation of Ehrlich and Holdren (1971).

capacity ($I_{m>E}$, see GFN, 2010) and hence decreases the levels of lifetime well-being of other current and future individuals. Also, the number of sentient beings on Earth is very high, so the welfare function reduces to the average of priority weighted levels of lifetime well-being (the population factor in the welfare function is close to 1, see section 4.6.3). If the environmental impact of the current moral agents decreases the lifetime well-being of other individuals, this average decreases and hence the welfare function decreases.

To reduce the environmental impact, the current population of moral agents has four options, referring to the four factors in the impact equation³: 1) decrease population size, 2) decrease lifetime well-being, 3) decrease consumption levels and 4) decrease the impact of resource consumption.

The most ethical approach for the first option consists in creating fair opportunities for a voluntary pregnancy restriction (by e.g. education for women and a good access to contraceptives and services for sexual and reproductive health). The second and third options refer to a lifestyle of voluntary simplicity. These two options imply that the moral agents should first cut on their resource use that does not strongly contribute to their lifetime well-being. Decreasing the consumption for luxury needs is a good starting point, because luxury needs (such as resource intensive social status symbols) are defined by the fact that society can create new circumstances where those needs no longer need to be satisfied in order to have an increase in well-being (see section 6.5). The fourth option can be done by technological innovations and scientific research.

We also observe that on planet Earth, the current population of moral agents has a fertility rate higher than the replacement level of roughly 2,1 children per female human. If the fertility rate remains that high, the future populations of moral agents will show an exponential growth. That is again unsustainable. As a consequence, future populations of moral agents and amoral sentient beings will receive much lower levels

³ Note that lowering the well-being of future people is a kind of impersonal harm. The current moral agents have two options: reduce or not reduce their environmental impact. If they do not reduce the impact, the future will contain a population A that will have low levels of well-being. If on the other hand the current moral agents reduce their impact, the future will contain a *different* population B, where the individuals have a higher lifetime well-being compared to the individuals in population A. Population B is different (the people in B are not the same people as in A), because the choices that the current moral agents make influence who will be born in the future. Suppose that the lifetime well-being in population A is still higher than 0, i.e. the lives in this population are still worth living. Nevertheless, population A suffers a harm, but this harm is impersonal because the alternative (when the current moral agents reduce their impact) would be that the future population A would not even exist. The harm consists in the lowering of the welfare function. The forefinger uses the welfare function, and hence looks at impersonal harms.

of lifetime well-being. That lowers the welfare function.⁴ In the end, the fertility rate of the population of moral agents should drop to the replacement level.

-**The middle finger:** stop the consumption for luxury needs. Luxury needs not only generate a high environmental impact, but they also result in a violation of the basic right (as was discussed in section 6.5): we have to avoid the use of non-sentient living beings for luxury needs.

-**The ring finger:** a decrease in environmental impact decreases the loss of biodiversity. This is another important reason why current moral agents should avoid overconsumption and overpopulation. The ring finger also says that even if the current planet Earth is overpopulated by overconsuming moral agents, everyone is allowed to procreate.

-**The little finger:** moral agents are allowed to be a bit partial, but should refrain from causing more harm by overconsuming resources. The little finger cannot justify the current levels of resource consumption.

-**The thumb:** every moral agent should give the good example by following a universalizable rule. For example having more than two children should be avoided as long as the fertility rate of the population of moral agents is higher than the replacement level. A moral agent cannot find a universalizable rule that allows for having more than two children in the current situation where the fertility rate is too high.⁵ The moral agent is allowed to follow a rule like “Everyone may have as many children as one likes, as long as the fertility rate is not above replacement level.” This is comparable to the universalizable rule “Everyone may take a train that one prefers, as long as there is some place available on the train.”

12.1.3 Intermezzo: maps of the moral landscape

Four of the five fingers can be expressed in a mathematical equation of the moral weight: $M=W+R+B$. The first term W is the welfare function of the forefinger (see section 4.6.3). The second term R represents the violations of the basic right of the middle finger, as well as the tolerated partiality principle of the little finger (see sections 6.5 and 6.6.2). The third term B corresponds with the value of biodiversity. This moral

⁴ Note that – for large populations – the population factor in the welfare function is close to 1 (see section 4.6.3). If – for smaller populations – this factor would be close to N/N_0 , an increase in the population size might increase the welfare function.

⁵ To decrease the environmental impact, I suggest that for a period of time the fertility rate of the population of moral agents should drop below the fertility rate, such that the population size can decrease to a sufficiently low level. After that, the fertility rate should increase to the replacement level to reach a sustainable steady state.

weight combines different moral forces into a quantity that should be maximized, just as the standard model of physics combines different physical forces into a quantity that is extremized (a Lagrangian, see Weinberg, 1996).

The thumb principle should be applied to this moral weight: derive those universalized guiding rules that result in a best situation under universal compliance. We should not try to maximize the moral weight directly by our actions. Instead, we should follow those rules that, when those rules are followed by everyone who is capable of following them, maximize the moral weight.

The moral weight is a multidimensional function of controllable variables that can be influenced by moral agents when those moral agents select rules. The controllable variables are e.g. distributable goods and liberties. These controllable variables and universalized rules are embedded in the moral weight as follows. The moral weight can be written as $M = M(x(h(v(r))))$, i.e. the moral weight is a function of values x which represent e.g. the levels of lifetime well-being, the strength of basic rights violations and the amount of biodiversity. The values x are functions of the world histories h (as seen in section 4.6). These world histories are dependent on the controllable variables v . Finally, these controllable variables are functions of the universalized rules r .

It is the rules that we (moral agents) have to select. So we first start with selecting a rule. Then we derive what distribution of controllable variables we would get if the rule is universalized. Hence, these universalized rules determine the distribution of the controllable variables. Next, a chosen distribution of controllable variables generates a number of possible world histories (there can be more than one world history for a unique choice of controllable variables, because the world can contain probabilistic uncertainties). Taking expectation values over those possible world histories gives us the values lifetime well-being, basic rights violations and biodiversity.

This moral weight function can be represented as a moral landscape, with peaks and valleys⁶. The peaks correspond with the best situations (e.g. the best distribution of goods and liberties).

As an example, take the controllable variable v that corresponds with a behavior that is natural, normal and necessary. This variable might represent e.g. the level of predation or the level of procreation that moral agents can control. If predation is universally prohibited, if the selected universalized rule r is “stop predation always and everywhere”, then we move to one end in the moral landscape as the variable v goes to zero. Then the B -term becomes very low because a lot of biodiversity gets lost. This

⁶ The concept of “moral landscape” introduced by Harris (2010) might correspond with a moral weight that only contains the welfare function W in a simplified form that represents sum-utilitarianism instead of quasi-maximin prioritarianism.

decrease of the B -term outweighs the increase of the W - and R -terms, and hence the moral weight decreases under universal compliance of a predation prohibition rule. A duty to stop predation moves us to a valley on the moral landscape.

Also, the B -term might have another very special property to flatten the moral weight function in some areas, such that e.g. $M(v)=M(v')$ for two levels of predation v and v' within an interval Δv . This means that we do not have a duty to (but we are allowed to) decrease the level of predation from v' to v . The more natural, necessary and normal a behavior is, the wider the range might be of the interval Δv that corresponds with the level of that behavior. This means that these types of behavior have a wide range of permissibility: we are allowed to increase or decrease those levels within that range.

The moral landscape allows us to visualize what kinds of actions are obligatory, permissible and impermissible. Suppose that if you do not do something (e.g. you do not help or harm anyone), we are at a specific point in the moral landscape, on a mountainside. If you harm others, you push location downwards to the valley on the moral landscape. If you help others, the location moves upwards and climbs the mountain. My intuition says that no-one should move the location downwards. Moving downwards is impermissible. My intuition also says that in moving upwards, we are allowed to choose in what direction the location can move upwards. This is the little finger principle of tolerated partiality (section 5.1). Any upward moving direction is permissible.

But sometimes some upward moving direction might be obligatory instead of merely permissible. Sometimes we should take a certain preferred path to climb up the mountain. Sometimes we have a duty to help others in a preferable way. When are you obligated to help in a specific way, even when you do not want to help in that way? This question can be answered by the following procedure. If someone forces you to help in that way, s/he uses you as merely a means (you have to do something that you do not want, and your presence is required to help others), and that means that the R -term adds a negative amount to the total moral weight. All else equal, this would result in a downward movement on the moral landscape. But by helping others, the W -term adds a positive amount, which means an increase of the total moral weight and an upward movement. If this W -term outcompetes the R -term (if the increase in the W -term is larger than the absolute value of the change in the R -term), it means that the increase in welfare trumps the violation of your basic right not to be used as merely a means. If that would be the case when you are forced to help, it implies that you have an obligation to help in that way.

We also have to consider the problem that the moral landscape is not unique and objective (impersonal): each moral agent might have his/her own preference for the parameters in the moral weight. These parameters correspond with e.g. the level of risk aversion or the relative strength of the R -term. Not everyone needs to have the same

levels of risk aversion, need for efficiency, estimates of well-being, or intuitive balancing choices between the strengths of different principles. Also the choice of welfare function can be different among moral agents. Therefore, each moral agent might look at his/her own map of the moral landscape. Those different maps might indicate different locations of the peaks and valleys. The highest peak on moral agent *a*'s map might also be much higher than the highest peak on *b*'s map. No map is the unique and objective one. No moral agent can be a dictator about e.g. the level of risk aversion that one should have.

There is a procedure that allows for democratic assessments between the preferences of all moral agents (section 4.6.11). First, all moral agents calculate their own moral weights and generate their own maps of the moral landscape. Second, for each moral agent a weighted moral weight is calculated: the welfare function M^a of moral agent *a* is divided by its maximum value M_{max}^a . If we now look at the moral landscapes representing the weighted moral weights according to moral agents *a* and *b*, we see that the highest peak on *a*'s map is as high as the highest peak on *b*'s map. In other words, the optimal situation according to moral agent *a* is as valuable as the optimal situation according to *b*. In the third and final step, a democratic average of all weighted welfare functions of all moral agents is taken:

$$\bar{M} = \frac{1}{N_a} \sum_{a=1}^{N_a} \frac{M^a}{M_{max}^a}.$$

All preferences of all moral agents (i.e. all persons who are able to do the exercise to derive what should be done) can and should be taken into account equally (democratically), making the theory more objective (impersonal).

This democratic procedure removes an important arbitrariness of the theory: one could object that my selection of parameters (e.g. the parameters that measures my risk aversion or my estimate of the strength of the basic right) is arbitrary. Why not take another level of risk aversion? The democratic procedure implies that everyone's preferences should be taken into account. This also means that e.g. the preferences of a utilitarian (i.e. no risk, loss and uncertainty aversion behind a veil of ignorance, no strengths for the mere means principle and the tolerated choice principle, and no value of biodiversity) can be included. Hence, it is possible to include the moral choices of someone who wants to delete some fingers (e.g. delete the deontological principle of the middle finger) and simplify the remaining fingers (e.g. simplify the forefinger principle).⁷

⁷ Deleting the thumb principle without opening the door for huge levels of arbitrariness, might be impossible. At least some version of a universalization principle should be preserved.

Constructing a huge number of maps (generated by weighted moral weights of all moral agents), each map representing a vast, multidimensional moral landscape, and taking the democratic average of those maps, will be very complicated in real life. Therefore, in daily life it is better to work with simpler rules of thumb to roughly approximate the location of mountainous areas.⁸ And we should set priorities to act against the greatest forms of injustice, such as poverty and animal abuse. It is clear that those forms of injustice are really far away from any peak. Even if we can't determine the location of the highest peak on the moral landscape, we do know that we better move to a mountainous area instead of remaining in a low area.

12.2 A second moral hand of meta-ethics

After removing the arbitrariness of the choice of parameters, only one important arbitrariness remains: why not take other moral fingers, why not include more principles, why not add more terms to the moral weight?

Perhaps such inclusions are possible and are allowed, under one important condition: they should be based on a coherent set of strong, shared moral intuitions which are translated in clear, universalized ethical principles. And that is not easy. Including a sixth principle to the moral hand might be incoherent with background theories (e.g. some scientific facts), moving the new theory away from a wide reflective equilibrium and turning this sixth principle into a moral illusion.

In other words: the thumb principle of universalization does not only apply to the choice of actions or behavioral rules ("If I am allowed to do something or follow a rule, then so are you."), but also to the choice of ethical systems ("If I am allowed to construct and follow an ethical system, then so are you."). The other four fingers of the moral hand strongly restrict what kind of action or rule we are allowed to follow. Similarly, there are four requirements that place strong constraints on the kind of ethical system we are allowed to construct: 1) internal (logical) consistency, 2) non-arbitrariness (simplicity), 3) clarity of formulations and 4) coherence with multiple strong intuitions.

Hence, there are two moral hands: the above mentioned moral hand with five moral principles of *normative ethics*, and a second moral hand of *meta-ethics*.

⁸ Again, we can compare this with physics. Solving the complete Lagrangian of the universe will be very difficult. This impractical difficulty does not imply that the standard model is wrong. Instead, physicists use approximations (e.g. mean field theory) to simplify things.

-**The thumb:** the principle of universalism. You must construct and follow the kind of ethical system that everyone who is capable must construct and follow in all similar ways. You may construct and follow the kind of ethical system that everyone who is capable may construct and follow in similar ways. An ethical system is a set of universalized ethical principles, applicable to all (real and hypothetical) situations.

-**The forefinger:** compatibility with the most important input data (the strongest and most shared moral intuitions) is important. These input data include intuitions derived from thought experiments and analogies of hypothetical situations. The input data also includes background theories such as reliable scientific knowledge (for example about biological species and evolution⁹, well-being and consciousness¹⁰,...). Non-scientific (e.g. pseudoscientific or religious) ideas should be excluded from the input data. With this condition, the ethical system will be in line with science.

-**The middle finger:** internal (logical) consistency is important.

-**The ring finger:** clarity (accuracy) of concepts and principles is important.

-**The little finger:** simplicity (parsimony, no artificial, ad hoc constructions, no arbitrariness) is important (cfr. Occam's razor).

Note that the fingers of the second moral hand correspond with epistemic virtues of scientific research (Kuhn, 1970).

We can compare this second moral hand with solving a crossword puzzle, where words correspond with universalized ethical principles.

The thumb says that one should always follow certain rules of the game to solve a crossword puzzle. For example the words should be existing words (i.e. the ethical principle should be correctly universalized to all similar situations).

The forefinger says that the words that one has to fill in should match the given descriptions of the words and the lengths of the words as good as possible. This implies that there should be no empty white boxes left: each white box should contain a letter. This also implies that a large crossword puzzle can be more coherent than a small puzzle. Hence, extending a 'moral crossword puzzle' by introducing new moral thought experiments and testing moral intuitions in exotic cases might increase coherence if these tests are successful.

The middle finger says that a white box cannot contain more than one letter. There should be no 'true' or 'hard' dilemmas in an ethical system, i.e. no judgment that something is both permissible and impermissible, all things considered.

The ring finger says that the words should be familiar words, not some strange, farfetched, incorrectly written or newly invented words.

⁹ The ethical system should respect the Darwinian fact that species do not have essences.

¹⁰ The ethical system should respect the scientific facts about personal identity, degrees of consciousness,...

The little finger says that a series of adjacent white spaces should preferably contain one word instead of more than one. E.g. if description “fruit” with ten letters can be either “strawberry” or “apple lemon”, the former solution is preferred. In general, we should avoid splitting rows and columns into smaller words (we should avoid introducing more ethical principles *ad hoc*). This principle also says that we should try to avoid making artificial (*ad hoc*) changes to a description in order to save a preferred word (e.g. if the description is “vegetable” and we want the word to be “apple”, we should avoid changing the description to “vegetable or fruit”).

Let us look at some examples how to apply the five fingers. As a first example, we can apply the thumb to the little finger: If you are allowed to define discrimination in a way that it refers to an arbitrary group (e.g. humans), then I may also pick an arbitrary group for my ethical rules, principles and definitions. My preferred group might exclude you, which you cannot want. Therefore, referring to arbitrary groups in rules, principles and definitions is not allowed.

Similarly, introducing a principle that prohibits gay sex and marriage violates the little finger, because we can all too easily ask the question: why gay sex? One could try to justify a prohibition on gay sex and marriage by referring to a principle of (sexual) purity. But then we get into trouble with the ring finger, because this concept of purity needs clarification (as I clarified the concept of e.g. naturalness). And then this clarified principle needs to be universally tested to see whether it is compatible with other important principles and intuitions. I doubt whether the notion of purity can be clarified and universalized to make it as clear and coherent as the five principles that I derived. The new anti-gay principle might easily get into conflict with scientific facts about gay people. Therefore it is unlikely that someone could simply include a principle (a sixth moral finger) that prohibits gay sex and marriage, because inconsistencies or strong levels of arbitrariness might appear after critical reflection and attempts to universalize this new principle.

The little finger (together with the forefinger) also has as a result that scientific facts can influence the ethical principles. Suppose for example that an ethical principle states that there is a sharp difference between individuals with moral status and those beings lacking moral status. This binary property of moral status cannot be matched non-arbitrarily to a natural property that comes in degrees. For example biological science can come to the conclusion that being human is a matter of degree (look at human ancestors and the possibility of hybrids, chimeras and genetically modified humans). As a result, matching an all-or-nothing interpretation of moral status with the natural property of being human cannot be done, except when one introduces an artificial or arbitrary cut-off point for being human. In the ethical system that I constructed in this dissertation, I always avoided matching a discrete (e.g. binary) property to a gradual property. In that sense, scientific facts (e.g. about gradual evolution, personal identity

or levels of complexity and mental capacities) had some influence in the construction of the ethical system.

As another example, we can apply the thumb to the forefinger, from which we can derive that an ethical system cannot be based on e.g. the Ten Commandments. If you are allowed to base your ethical system on ideas (e.g. the existence of a Christian God) that lack evidence, I am allowed to do so as well. So I am allowed to invent things that equally lack evidence. For example I can introduce another God with other commandments that will harm you, and you cannot want that. The belief in e.g. a Christian God is too arbitrary (it violates the little finger): the evidence for the existence of Krishna, Apollo or Zeus is as high as the evidence for the existence of God. In this sense, all theistic believers are in fact inconsistent atheists: they do not believe in all the other gods, they are not willing to have blind faith in another god, they are not open-minded towards another god. Such levels of arbitrariness and inconsistency are not tolerable in ethics. Therefore, there is no room for religious faith-based ethical principles in the construction of an ethical system.

The first moral hand of normative ethics required some intuitive balancing. For example the mere means principle of the long middle finger is stronger than the tolerated partiality principle of the small little finger. Similarly, the moral hand of meta-ethics requires an intuitive balancing between the different criteria for a good ethical theory. For example the consistency requirement of the middle finger is stronger than the simplicity requirement of the little finger.

The most important point is that the meta-ethical moral hand should set very high standards for a good ethical system. High standards are required to restrict the number of possible ethical systems, to avoid an 'anything goes' attitude towards ethics and to limit moral relativism. For example applying the meta-ethical thumb principle to the forefinger results in a completeness requirement: we should use analogies or thought experiments involving all possible situations. A good ethical theory has to be able to deal with all realistic as well as hypothetical situations. Even very hypothetical situations count: the better an ethical system is able to deal with all kinds of hypothetical situations, the better the system is.

Setting high standards is what I did in my derivation of a system of animal equality. I dealt with very hypothetical situations: what if insects or plants were sentient? What if a predator (a lion) acquired full blown moral agency? What if we bred and used mentally disabled humans as slaves? What if we could teleport and make (inexact) copies of persons? What if the side track in the trolley dilemma loops back to the main track? What if human-animal hybrids or chimeras would exist?

Using the ring finger of the meta-ethical hand, I tried to clarify concepts such as well-being, biodiversity, naturalness, the body and use as merely a means. In these

clarifications, I tried to avoid arbitrariness and artificiality in definitions. Also the use of mathematical expressions should be understood as an attempt to clarify principles.

By restricting the complete normative ethic to five principles, I also tried to make the theory as simple and parsimonious as possible, respecting the little finger of the meta-ethical hand. Utilitarians might complain that my theory remains too complex. Their theory is more economical, but it goes at the cost of violating the meta-ethical forefinger (it has a lower match with moral intuitions that many people share). So here we end up with an important trade-off between the forefinger and the little finger. As the forefinger is a little bit longer, my constructed ethical system tends to lean a bit more towards compatibility (with intuitions) than simplicity.

12.3 The impossible triangle of the meat eater

Meat eaters are often not aware of the inconsistency of their meat consumption. In fact, with the antidiscrimination principle of the thumb and the mere means principle of the middle finger (applied to humans), we can construct an instructive analogy between this inconsistency of meat consumption and the optical illusion of the impossible triangle.

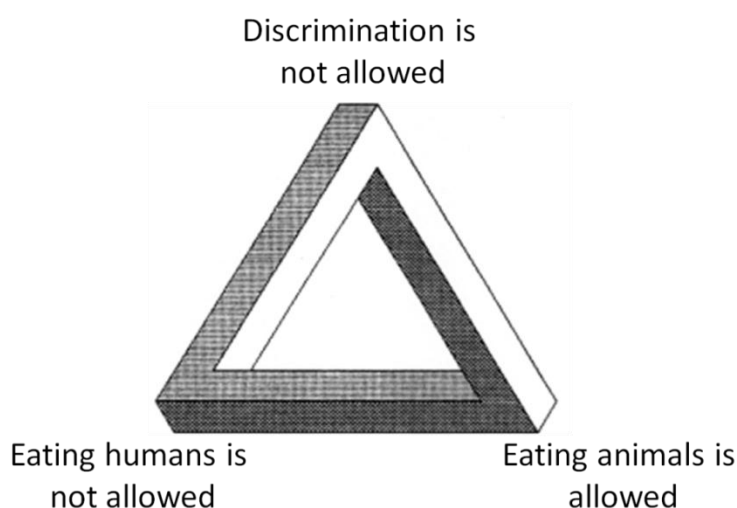


Figure 14: the impossible triangle of the meat eater

Most meat eaters share the same moral intuitions, that discrimination (partiality based on arbitrary, morally irrelevant criteria) is not allowed, that killing and eating humans is not allowed because all humans have the basic right not to be used as merely a means, and that eating animals is allowed, because we should respect our freedom of

choice (it's not nice to prohibit something to someone). Each of these principles can be represented by the corners of a triangle.

Now, all I'm asking from the meat eater is to zoom out. If you look at the above figure, and you only focus at one or two corners at a time, you will not see the contradiction. Antidiscrimination is compatible with not eating humans. But introducing the third principle turns the whole into a contradiction. The impossible triangle is an optical illusion: it cannot exist in reality. I have confidence that the meat eater is capable of noticing the contradiction when s/he zooms out and looks at the whole picture at once. Especially after presenting ten arguments about the moral (ir)relevance of species and sentience, as I have done in a previous chapter 8.4), it is clear that speciesism is discrimination.

So what can we do now, after we have accepted that the picture is an optical illusion? We can do one of two things. First, we can simply delete one of the principles; simply erase one of the three corners. Then we end up with a bar with two endpoints (two principles). This bar can exist in reality. Another possibility is to redraw the picture, turning one of the corners in the other direction.

The question I want to ask to the meat eater is: which corner do you want to turn? Do you want to allow cannibalism? Do you want to allow racism, sexism or other forms of discrimination? Or do you want to prohibit eating animals? I am convinced that most meat eaters would prefer the third option, simply because the other two are much less attractive. First, the intuition that we are allowed to eat meat is much less strong than the intuition that cannibalism or discrimination are not allowed. Adopting the third principle of freedom of choice is not so difficult. We could simply state that we should respect freedom of choice, except when the basic right is violated. So we are still allowed to eat and choose what we like, but we have to restrict our choices a little bit.

Second, from behind a veil of ignorance, we would not like to live in a world where cannibalism or discrimination is allowed. If the meat eater is rational and cares for his/her well-being, s/he would prefer a world where s/he could not enjoy the taste of meat above a world where s/he might be the victim of cannibalism or discrimination.

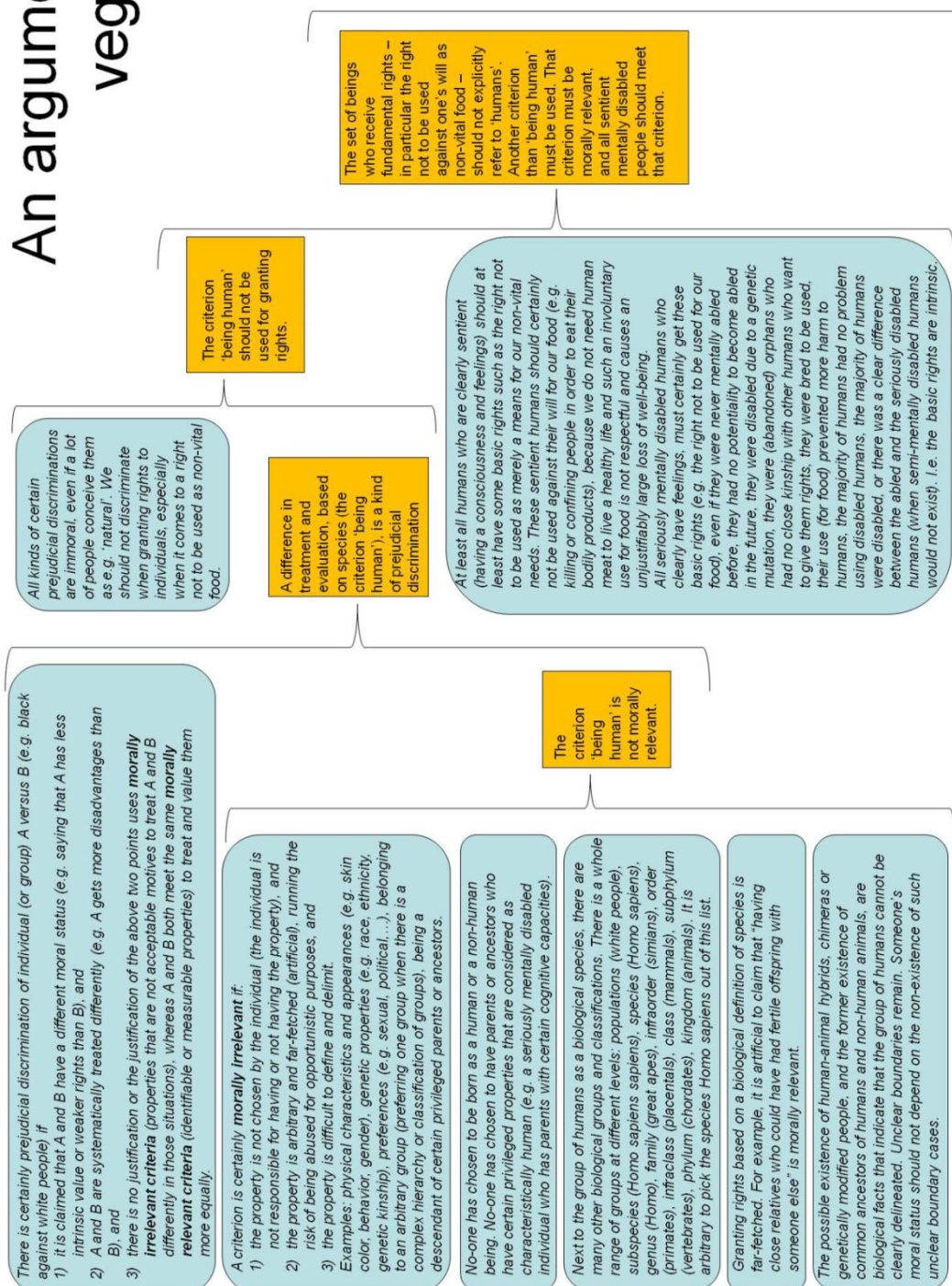
So we have two coherent arguments that indicate that it is best to turn the lower right corner of the impossible triangle. Or stated in another way: if we (I and a meat eater) put all our intuitions on the table, we restrict ourselves to those intuitions that we both share, we give those intuitions a ranking according to their strength, we translate those intuitions into universalized ethical principles, we construct a consistent ethical system with those principles whereby we systematically give priority to the strongest principles, then we end up with a system of animal equality. That system implies veganism.

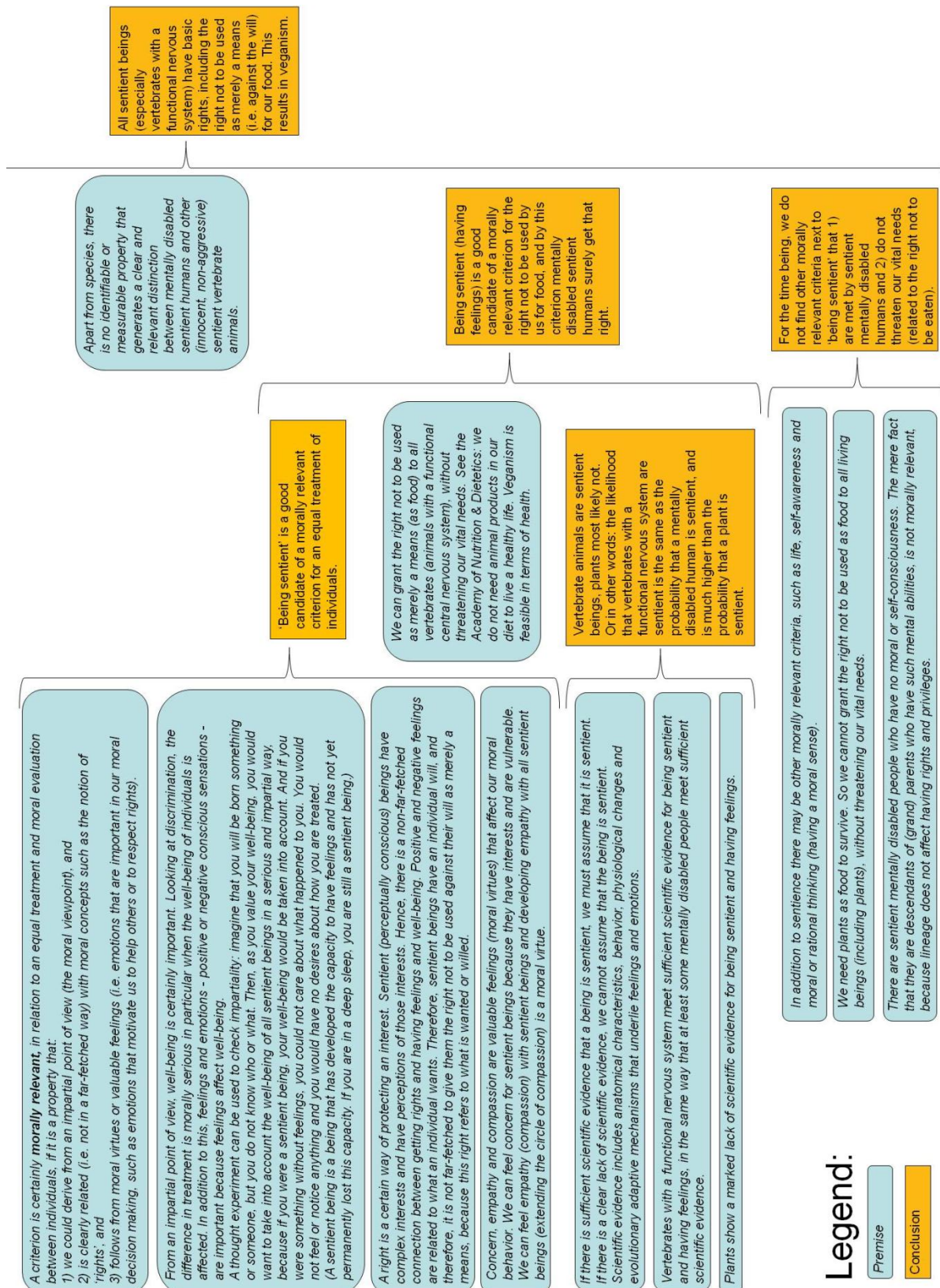
Veganism is a very simple rule of thumb: do not use animal products. From a political perspective, the technical implementation of veganism might be rather easy. Already today, a lot of animals are subject to welfare laws. These are the animals that are

sentient, according to scientists. So we already deal with sentience in the law. We do not have to change the set of beings that are subject to welfare laws, we only need to change the content of the rights that those beings have. Everyone who is now already subject to welfare laws (everyone who already has some welfare rights), should get the same rights as mentally disabled humans already have. Use your wildest imagination: what if we treat mentally disabled humans in the same way that we treat some animals in e.g. medical experiments or factory farms? If we should not treat those humans in those ways, then we should also abstain from treating in those ways everyone else who is subject to current welfare laws.

12.4 Argumentation scheme for veganism

An argument for veganism





Where to go from here? Questions for future research

Of course not all problems related to ethical consistency and animal equality are solved. How can we refine the argument about speciesism? How can we strengthen the method to detect moral illusions? How can we further clarify the basic ethical principles? These are the rather obvious questions for future research. So let us have a brief look at the less obvious but more intriguing questions.

I have constructed a pluralist, principlist ethical system of the moral hand, which contains five ethical principles.

1) Why this set of principles instead of another pluralist principlist system such the four principles of Beauchamp & Childress (2001), or the seven principles of Ross (1930)? I think that the five principles of the moral hand are in fact some reshuffling of the principles of Beauchamp, Childress and Ross, where the reshuffling is done such that the resulting five principles can be expressed in a more compact and yet clear way. I will not further elaborate on this issue here.

2) Is five too much or not enough? Consequentialist welfare ethicists would favor a simplified system with only one finger: the forefinger. Sum-utilitarians would furthermore simplify the forefinger, deleting the priority for the worst-off. Rule consequentialists might prefer two fingers: the thumb (which refers to the universalized rules) and the forefinger. On the other hand (no pun intended), some deontological libertarians might restrict their ethical system to the basic right principle of the middle finger. But perhaps – moral particularists might argue – five is not enough? Perhaps moral particularists are right: maybe moral judgments are like esthetical judgments of music. When you hear a piece of music, you automatically judge it to be good or bad, but can you capture all of your musical taste preferences and judgments in a small set of principles? Doing that would seem to be a miracle, discovering the magic potion of music. Perhaps ethics, like music, is far too rich to be expressed in five principles. Perhaps not even hundred principles will do.

3) What is the strength of the different principles? In other words: how long and strong are the five fingers? A sum-utilitarian is a pluralist who gives absolute strength to only one principle: the forefinger. The other fingers have zero length, they have zero value. A deontological libertarian gives absolute strength to the middle finger principle. But most of us are real pluralists: different principles have some non-zero strengths and fingers have some non-zero lengths. How do we balance the strengths of the different principles? In previous sections I have suggested a kind of democratic procedure to solve this problem: every moral agent has an equal vote to put forward his/her intuitive judgments about the relative strengths of the different principles. If most of us put greater weight to the mere means principle of the middle finger, we have to accept this outcome; we have to accept that the average middle finger is longer than the average forefinger. So, everyone has his or her moral hand, and we measure the average lengths of everyone's thumb, forefinger and so on to derive the 'average' or 'platonic' moral hand.

But there are two problems with this procedure.

First, how could we follow the democratic procedure in practice? Taking the intuitive judgments of all moral agents into account becomes as complicated as constructing a combustion engine using the standard model of elementary particle physics. But at least we can try to find some rough moral rules of thumb to guide us, just as we use the laws of thermodynamics.

A second, more intriguing problem is that studies in moral psychology clearly demonstrated that the intuitive strength of a moral principle depends on external circumstances. For example induced feelings of disgust (Schnall et al., 2008) and happiness (Valdesolo & DeSteno, 2006) can influence moral intuitions in the trolley dilemmas, making us more or less reluctant to sacrifice someone in order to save others. So the lengths of the fingers of the moral hand depend on whether or not we are disgusted, happy, tired, sniffed some oxytocine, saw a good movie,... It is like fingers can grow and change lengths at different speeds. Imagine that we could 'nudge' someone's moral intuitions, such that we can turn a utilitarian into a libertarian by tweaking his/her intuitions. What would this imply for our procedure to derive the strengths of the principles? Is there a 'neutral state' where our moral intuitions are not influenced by feelings of disgust or happiness?

4) Finally, let me pose the most intriguing question: where did it come from? I have constructed an ethical system, but what exactly did I do? Did I discover it or invented it? It is like the question what mathematicians do: are they discovering mathematical facts in a platonic world of mathematics, or are they rather inventing and constructing theories like engineers do? Are the moral fingers just some clever confabulations and rationalizations of a bunch of mysterious intuitions? Where do these intuitions come from in the first place? Do they have an evolutionary psychological explanation? Did

evolution really have some influence in how I defined the mere means principle or the notions of well-being and biodiversity?

The more I think about this, the more mysterious it all seems. And yet, I hope that I have made some progress.

Nederlandstalige samenvatting: dieren, dilemma's en discriminatie¹

Inleiding

De discussie over de morele status van niet-menselijke dieren², en het gebruik ervan voor voeding, kleding, amusement of onderzoek, heeft een lange geschiedenis die teruggaat tot de Griekse oudheid (Pythagoras, Plutarchus). Doorheen de eeuwen stak de problematiek af en toe opnieuw de kop op (Leonardo Da Vinci; Jeremy Bentham; Oswald, 1791; Ritson, 1802; Salt, 1892), maar een echte doorbraak voor de academische dierenethiek kwam er in 1971, toen Richard Ryder de term *speciesisme* (of *soortisme*) lanceerde: een discriminatie van dieren ten opzichte van mensen, naar analogie van racisme en seksisme (Godlovitch & Harris, 1971; Ryder, 1975). De jaren '70 en '80 werden gekenmerkt door toepassing van verscheidene rationele theorieën in de normatieve ethiek (voornamelijk utilitarisme en deontologische ethiek) op het dierenvraagstuk (Singer, 1975; Clark, 1977; Regan, 1983). In de jaren '90 kwam er, vanuit postmodernistische en feministische hoek, kritiek op deze 'kille' rationele aanpak (Adams, 1995). Een nieuw pleidooi voor vegetarisme steunde op bv. een zorgethiek (Adams & Donovan, 1996) of een deugdenethiek (Hursthouse, 2000). Rond de eeuwwisseling nam de discussie een nieuwe wending, naar de (sociale) psychologie. Vele dierenethici beschouwen de argumentatie voor dierenrechten en veganisme als afgerond, maar stellen vast dat er meer psychologie dan ethiek schuilt achter ons gebruik van dieren (Serpell, 1996; Allen et al. 2000; Joy 2002, 2009; Herzog 2010). De

¹ Deze samenvatting is gebaseerd op Bruers (2013b)

² Ik zal "niet-menselijk dier" meestal afkorten tot "dier". Met "vleeseter" bedoel ik een mens die minstens af en toe vlees eet. "Vegetarisme" en "veganisme" zal ik vaak gebruiken als synoniemen, hoewel de implicaties van mijn dierenethiek veganisme zijn.

vraag dient gesteld: Waarom zijn zo weinig mensen overtuigd door logisch consistente, rationele argumentaties?

Met dit onderzoek keer ik terug naar de filosofie van de jaren '70 en '80, en diens zoektocht naar een consistente dierenethiek. Na grondige beschouwing van 40 jaar dierenrechtenethiek zet ik hier een argumentatie neer op basis van morele intuïties die velen onder ons delen. Ik presenteer een uitgewerkte dierenethiek aan de hand van een metafoor: *de morele hand*. Vijf vingers staan symbool voor vijf ethische basisprincipes waarop deze ethiek stoelt. De volgende vijf onderdelen van deze Nederlandstalige samenvatting van het doctoraatsonderzoek gaan over universalisme (de duim), het belang van welzijn (de wijsvinger), het basisrecht om niet als louter middel te dienen (de middelvinger), de waarde van biodiversiteit (de ringvinger), en een toelaatbare vorm van partijdigheid (de pink). Zoals men de duim moet plaatsen tegenover de overige vingers om een voorwerp te kunnen grijpen, zo ook moet men de universalistische duimregel toepassen op de vier andere ethische principes om een ethisch probleem te kunnen vatten. Een ethisch systeem wordt coherent (intern samenhangend) wanneer de universalistische principes consistent zijn. Het opbouwen van een coherent ethisch systeem is een methode om *morele illusies* te onderzoeken. Dergelijke illusies zijn – naar analogie van optische illusies – hardnekkige intuïtieve oordelen die 'fout' zijn in de zin dat ze in tegenstrijd zijn met een coherent systeem van universalistische ethische principes die gebaseerd zijn op andere, sterkere morele intuïties.

Het eerste deel (het universalismeprincipe - de duim) bespreekt de methode van coherentie en toont aan dat speciesisme een morele illusie is. Uit het universalismeprincipe volgt een formeel principe van gelijkwaardigheid in termen van onpartijdigheid en antidiscriminatie: behandel gelijken gelijk in gelijkaardige situaties.

De twee daaropvolgende hoofdstukken kaarten de twee grootste stromingen binnen de dierenethiek aan: een consequentialistische ethiek (de wijsvinger), die kijkt naar de gevolgen voor het welzijn van alle voelende wezens; en een deontologische ethiek (de middelvinger), die vertrekt van de plicht om een basisrecht te respecteren. De wijsvinger- en middelvingerprincipes genereren zo twee verschillende vormen van gelijkwaardigheid.

Maar deze twee stromingen kampen met een ernstig onderschat probleem binnen de dierenethiek: het predatieprobleem. Uit persoonlijke contacten met dierenactivisten en –ethici vermoed ik dat de morele intuïties met betrekking tot dit probleem een sterke consensus vertonen: de meeste dierenethici zijn erg tolerant tegenover het gedrag van roofdieren. Daar tegenover staat de intuïtie dat iemand doden voor een gedwongen orgaantransplantatie niet mag. Nochtans wordt in beide gevallen (predatie en transplantatie) een voelend wezen (prooidier, orgaanslachtoffer) zonder diens toestemming gedood en wordt diens lichaamsweefsel (spieren, organen) opgenomen in een ander voelend wezen (roofdier, ziekenhuispatiënt) met als doel te blijven leven.

Gedwongen predatie mag wel, gedwongen orgaantransplantatie niet. Is dit intuïtieve onderscheid tussen predatie en transplantatie een morele illusie? Als we die intuïtie willen respecteren en niet afdoen als morele illusie, zijn we genooddaakt om een nieuw universalistisch ethisch principe in te voeren. Dit is wat de ringvinger biedt. Het vierde hoofdstuk presenteert een drietal mogelijke oplossingen voor het predatieprobleem, waarbij het ringvingerprincipe, gebaseerd op de waarde van biodiversiteit, mijn persoonlijke voorkeur geniet. Uit het ringvingerprincipe volgt een nieuw gelijkwaardigheidsprincipe: iedereen heeft een gelijk recht op een gedrag dat tegelijk natuurlijk, normaal en noodzakelijk is (gedrag dat sterk bijdraagt aan de biodiversiteit). Als een zebra mag eten om te overleven, mag een leeuw dat ook.

Bij predatie komt er nog een probleem kijken dat in het laatste hoofdstuk besproken wordt: hoewel we prooien niet moeten beschermen, voelen we toch een sterke plicht om mensen te redden van roofdieren. Is het redden van een mens, en niet een zebra, een vorm van discriminatie, of is een partijdigheid voor een mens in dergelijke situaties toch nog verzoenbaar met antispeciesisme? In hoeverre mogen we nog partijdig zijn tegenover onze dierbaren? Om een te veeleisende ethiek te voorkomen, zal de vijfde vinger – de kleine pink die een klein beetje kan afwijken van de richting van de andere vingers – aangeven welke kleine vorm van partijdigheid en afwijking van de andere principes nog aanvaardbaar zijn. Uit dit partijdigheidsprincipe van de pink kunnen we een subtiele vorm van gelijkwaardigheid afleiden. Samen met de vier vormen van gelijkwaardigheid die volgen uit de andere vingers, schets dit een genuanceerd beeld van de precieze betekenis van dierengelijkheid. Werken de vijf morele vingers als morele hand samen, dan komen we zo tot een consistente ethiek van dierengelijkheid. Op basis van de vijf principes van de morele hand kunnen we concluderen dat veganisme een morele plicht is.

1. De duim: het universalismeprincipe, coherentie en morele illusies

Universalisme is belangrijk omwille van twee redenen. Ten eerste vormt universalisme samen met consistentie de basis van *coherentie* waarbij verschillende principes of intuïties elkaar wederzijds ondersteunen. Deze coherentie ligt ten grondslag aan de methode om morele illusies op te sporen. Ten tweede kent universalisme ook een specifieke vorm: het *antidiscriminatieprincipe*, waarbij men alle gelijkwaardige individuen gelijk moet behandelen. Met vijf coherente argumenten zal ik trachten aan te tonen dat het mens-dier waardeonderscheid (speciesisme) een vorm van

discriminatie is, en met nog extra coherente argumenten zal ik nog een stap verder gaan en trachten aan te tonen dat dit speciesisme gebaseerd is op een morele illusie. Net zoals bij optische illusies kan dat illusoire karakter van speciesisme de hardnekkigheid van speciesistische oordelen in onze cultuur verklaren. Net zoals bij optische illusies genereren morele illusies hardnekkige oordelen (bv. racistische of speciesistische oordelen) die niet compatibel zijn met andere sterke oordelen (bv. oordelen over rechtvaardigheid en onpartijdigheid).

1.1. Coherentie als universalisme plus consistentie

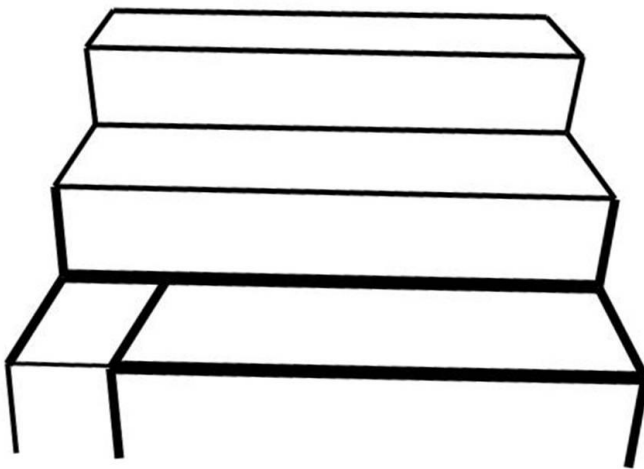
De coherentie van een ethische theorie werkt zoals de samenhang van een kruiswoordraadsel, waarin kruisende woorden elkaars geldigheid bevestigen. De witte vakjes in een kruiswoordraadsel komen overeen met verschillende specifieke situaties of gezichtspunten. Een letter in een vakje is dan het analogon van een moreel oordeel over die specifieke situatie vanuit dat specifieke gezichtspunt. Het construeren van een coherent ethisch systeem is dan analoog aan het invullen van een kruiswoordraadsel. Om een coherent systeem te bekomen, zijn twee elementen belangrijk: universalisme en consistentie.

Universalisme wil zeggen dat de letters in rijen en kolommen van aangrenzende witte vakjes van het kruiswoordraadsel een onderling verband moeten hebben. Het zijn niet zomaar willekeurige, onafhankelijke letters: ze vormen samen een woord. Zo moeten morele oordelen in alle gelijkaardige situaties (alle aangrenzende vakjes) worden verklaard volgens een onwillekeurig universeel ethisch principe. Ethisch universalisme is het vertalen van sterke morele intuïties naar universalistische ethische principes, universeel geldend voor alle gelijkaardige situaties. Dit universalisme legt dus een eerste strenge voorwaarde op bij de constructie van een ethisch systeem, net zoals er een strenge voorwaarde opgelegd wordt bij het oplossen van een kruiswoordraadsel.

Universalisme is echter niet voldoende, ook consistentie is belangrijk. Zoals men geen verschillende letters mag invullen in éénzelfde witte vakje van een kruiswoordraadsel, mag men geen twee, of meer, tegenstrijdige morele oordelen of ethische principes hebben. Consistentie levert een tweede strenge voorwaarde, en enkel door het samengaan van beide voorwaarden zullen de verschillende woorden in een kruiswoordraadsel elkaar wederzijds ondersteunen en bevestigen. Als er tot slot geen lege witte vakjes meer zijn, en als er geen situaties zijn met onbesliste morele oordelen, dan is het coherente systeem volledig.

1.2. Morele illusies

Morele illusies (hardnekkige, incoherente morele intuïties) kunnen nu worden onderzocht aan de hand van coherentie. We kunnen hierbij gebruik maken van de analogie met optische illusies, in het bijzonder de Müller-Lyer-illusie. De meeste mensen die naar de Müller-Lyer-figuur kijken, hebben een spontaan oordeel dat de lengte van één lijnstuk langer is dan dat van een ander. En zelfs als ze inzien dat de lijnstukken toch even lang zijn, blijft de illusie hardnekkig aanwezig. De volgende tekening van een trap bevat een Müller-Lyer-illusie: de onderste dikke horizontale lijn lijkt korter dan de dikke lijn erboven, maar op papier zijn ze even lang.



Figuur 15: de Müller-Lyer-illusie in de tekening van een trap. Let op de lengte van de dikke horizontale lijnen.

Hoe tonen we aan dat de Müller-Lyer-illusie daadwerkelijk een illusie is? Ik wil in de ethiek niet verwijzen naar een objectieve morele realiteit, want het bestaan van zulke realiteit kent veel filosofische onenigheid. Zo ook wil ik de optische illusie kunnen aantonen zonder een beroep te doen op een objectieve fysische realiteit. Door de methode van het coherentisme kunnen we met enkel onze eigen intuïties reeds aantonen dat sommige van die intuïties onbetrouwbare illusies zijn. Net zoals we onze meetkundige intuïties kunnen vertalen naar universalistische axioma's, zo kunnen we ook in de ethiek morele intuïties vertalen naar universalistische basisprincipes. Om de Müller-Lyer-illusie aan te tonen, maak ik gebruik van meetkundige intuïties die vertaald naar universalistische basisprincipes luiden als:

1. *Translatie-invariantie*: de lengte van een meetlat blijft altijd dezelfde als men de meetlat verschuift in alle mogelijke richtingen.
2. *Contextonafhankelijkheid*: de lengte van een lijnstuk wordt nooit bepaald door de aanwezigheid van eender welke andere geometrische figuren rondom het lijnstuk.

Deze basisprincipes zijn gebaseerd op onze intuïties: spontane oordelen die we niet verder kunnen bewijzen. Zo bekomen we twee, op intuïties gebaseerde, argumenten om de optische illusie aan te tonen; het verschuiven van een meetlat, en het afdekken van de kleine schuine streepjes op de uiteinden van de horizontale lijnstukken. De twee meetkundige intuïties van translatie-invariantie en contextonafhankelijkheid wegen sterker door dan de ene intuïtie die zegt dat de horizontale lijnstukken niet even lang lijken. Het eenvoudigste wat we nu kunnen doen, is erkennen dat deze laatste intuïtie fout is: het is een optische illusie. Dat aanvaarden is eenvoudiger dan een meetkundig systeem creëren zonder translatie-invariantie of contextonafhankelijkheid.

Om het plaatje nog coherenter te maken, kunnen we kijken naar een psychologische verklaring van de Müller-Lyer-illusie. Zoals de tekening van de trap reeds doet vermoeden, ontstaat de illusie doordat onze hersenen tweedimensionale beelden interpreteren als driedimensionale objecten. Onze hersenen zijn getraind om lengtes in drie dimensies in te schatten, en daardoor maken ze een perspectiefaanpassing als we naar een tweedimensionale figuur kijken. Deze aanpassing werkt zoals een intuïtieve vuistregel of heuristiek (Kahneman, 2003) die meestal goed werkt, maar soms een foutief oordeel genereert: de optische illusie. De heuristiek waarmee we dit oordeel vormen, ontstaat pas na een leerproces, nadat de hersenen hebben leren omgaan met dieptezicht en perspectief. In gebieden waar geen trappen of andere rechthoekige objecten voorkomen, zijn inheemse mensen minder vatbaar voor de illusie. De Müller-Lyer-illusie blijkt zowaar aangeleerd (Ahluwalia, 1978).

1.3. De universalistische duimregel en antidiscriminatie

De ‘morele duim’ die het universalisme symboliseert, is cruciaal in een ethiek, want zonder een opponeerbare duim kunnen we moeilijk een probleem aanpakken. Een versie van de universalistische regel luidt: je moet de regels volgen die alle moreel bewuste wezens die ertoe in staat zijn zouden moeten volgen in alle gelijkaardige situaties; je mag enkel de regels volgen die alle moreel bewuste wezens die ertoe in staat zijn mogen volgen in alle gelijkaardige situaties. ‘Alle gelijkaardige situaties’ betekent ook: tegenover alle gelijkaardige individuen. Als men gelijkwaardige individuen ongelijk behandelt, kan er sprake zijn van schadelijke discriminatie. De universalistische duimregel genereert dus een formeel principe van gelijkwaardigheid in termen van onpartijdigheid en antidiscriminatie. Dat is een formeel principe, omdat het nog niet zegt hoe we individuen moeten behandelen.

Er is al zeker sprake van schadelijke discriminatie tussen twee partijen A en B (bv. man en vrouw, blanke en zwarte) indien er aan de volgende vier voorwaarden voldaan is:

- 1) A en B worden verschillend moreel gewaardeerd (bv. het toekennen van een hogere morele status aan A dan aan B), én
- 2) A en B worden systematisch verschillend behandeld (bv. het ontzeggen van rechten van B ten opzichte van A) , én
- 3) A en B voldoen beide aan dezelfde moreel relevante criteria om ze gelijk te waarderen en te behandelen, én
- 4) men tracht het verschil in behandeling te rechtvaardigen door te verwijzen naar eigenschappen die eigenlijk moreel irrelevant zijn en geen aanvaardbaar motief vormen om A en B in die situatie verschillend te behandelen.

Dierenethici (Ryder, 1975; Singer, 1975) argumenteren dat speciesisme voldoet aan deze vier voorwaarden, en dus een vorm van schadelijke discriminatie is: de soortgrens is geen moreel relevant criterium, het hebben van gevoelens is dat wel. Maar mijn argumentatie gaat nog een stap verder. Speciesisme vertoont, bij nader onderzoek, immers een zeer sterke gelijkenis met de Müller-Lyer-illusie.

1.4. Speciesisme als morele illusie

De analogie tussen de meetkunde als axiomatisch systeem gebaseerd op geometrische intuïties en de ethiek als axiomatisch systeem gebaseerd op morele intuïties, kunnen we doortrekken naar de dierenethiek. De meeste mensen hebben een intuïtief oordeel dat één lijnstuk in de Müller-Lyer-figuur langer is dan het andere, maar deze intuïtie is in strijd met een coherent systeem van andere, sterkere intuïties. De meeste mensen hebben ook de intuïtie dat mensen meer morele waarde hebben dan dieren. Wat ik zal trachten aan te tonen, is dat deze intuïtie (het dier-mens waardeonderscheid) in strijd is met een coherent systeem van andere, sterkere morele intuïties. De Müller-Lyer-illusie is dus analoog aan de morele illusie die speciesisme veroorzaakt. We zouden kunnen stellen dat het onderste dikke horizontale lijnstuk in de Müller-Lyer-illusie symbool staat voor de intrinsieke waarde of morele status van een dier, en het bovenste dikke lijnstuk symbool staat voor de waarde van een mens. Des te langer het lijnstuk is, des te hoger de morele status. De schuine streepjes staan dan symbool voor de moreel irrelevante eigenschappen (cfr. het vierde van bovenstaande criteria voor discriminatie). Die eigenschappen creëren de illusie waardoor het lijkt alsof mensen meer waard zijn (een 'langer lijnstuk' hebben).

Ik zal eerst vijf argumenten geven die suggereren dat het criterium *Homo sapiens* moreel irrelevant is.³ Deze argumenten zijn analoog aan de contextonafhankelijkheid in de Müller-Lyer-illusie. De invloed van de schuine streepjes op de lengte van de dikke horizontale lijnstukken zou zeer arbitrair, gecompliceerd en vergezocht zijn. Waarom zou het lijnstuk inkrimpen als de streepjes precies op de uiteinden staan en pijlpunten vormen die naar buiten wijzen? Hoe zou de lengte beïnvloed kunnen worden door de hoek tussen de kleine streepjes? Ook het criterium ‘mens zijn’ is een dergelijke arbitraire, gecompliceerde en vergezochte regel, zoals blijkt uit volgende vijf argumenten.

1) Wie ben ik? Ikzelf behoor tot de populatie van *blanken*, de ondersoort *Homo sapiens sapiens*, de soort *Homo sapiens*, het genus *Homo*, de familie van *mensachtigen*, de superfamilie van *mensapen*, de infraorde van *smalneusapen*, de orde van *primaten*, de klasse van *zoogdieren*, de onderstam van *gewervelden*, de stam van *chordadieren*, en het rijk van de *dieren*. Het is arbitrair om zonder rechtvaardiging de derde rangorde (de soort) te nemen als ‘ingroup’, als morele gemeenschap. Waarom de derde en niet bijvoorbeeld de vierde?

2) Veel mensen trachten het waardeonderscheid tussen mens en dier te rechtvaardigen door te verwijzen naar hogere mentale vermogens (bv. rationaliteit). Maar als ze dan geconfronteerd worden met het argument van de mentaal gehandicapten (mensen die behoren tot de morele gemeenschap maar nooit die hogere mentale vermogens hadden of zullen hebben), antwoorden ze dat die gehandicapten nog wel tot dezelfde soort behoren als rationele mensen. Maar het biologische soortbegrip is erg abstract en gecompliceerd. Eén van de vele definities van het begrip soort verwijst naar het kunnen krijgen van vruchtbare nakomelingen. Uiteindelijk lijkt het wel alsof men de morele status van mentaal gehandicapten tracht te funderen aan de hand van een regel die zegt dat als een wezen (bv. een mentaal gehandicapte) naaste familieleden (bv. ouders) heeft die vruchtbare nakomelingen (hadden) kunnen krijgen met een verzameling van wezens die bepaalde mentale vermogens (bv. rationaliteit) hebben, dat dan dat wezen zelf ook behoort tot de morele gemeenschap. (De verwijzing naar naaste familieleden is om rekening te houden met de mogelijkheid dat het wezen zelf door bv. een ziekte onvruchtbaar is.) Maar zulke regel is bijzonder vergezocht en artificieel. En het is oneerlijk om te zeggen dat een individu rechten heeft omdat diens ouders reproductief succesvol waren en in staat waren iets te doen met anderen. Het is

³ Voor alle duidelijkheid: dit criterium verwijst naar een biologische eigenschap. Verwijzingen naar mentale vermogens (rationeel denkvermogen,...), potenties (het kunnen ontwikkelen van mentale vermogens), familiebanden of sociale relaties zijn irrelevant indien men bv. aan ongeneesbaar diep mentaal gehandicapte weeskinderen een hoge morele status wil toekennen. Dit is het ‘argument van de randgevallen’ (‘argument of marginal cases’; Dombrowski, 1997).

onrechtvaardig om het geven van rechten op een dergelijke manier af te laten hangen van iemands familieleden. Dat bepaalde wezens “pech hebben de verkeerde familieleden te hebben”, mag geen rol spelen bij de keuze om die wezens al dan niet rechten te geven.

In verband met deze soortdefinitie kunnen we ook kijken naar zogenaamde ringsoorten, zoals de *Larus* meeuwen of de *Ensatina* salamanders. Dergelijke ringsoorten bestaan uit een reeks populaties, waarbij iemand van populatie A vruchtbare nakomelingen kan krijgen met iemand van populatie B, en hetzelfde voor B met C, en C met D, maar tussen populaties D en A lukt het niet meer. Net zoals ringpopulaties een verwantschap hebben volgens een ruimtelijke spreiding, hebben alle soorten een gelijkaardige relatie doorheen de tijd. Kijken we naar de fylogenetische stamboom, dan zien we dat een moderne *Homo sapiens* biologisch in staat is om vruchtbare nakomelingen te krijgen met voorouders. En die voorouders met nog vroegere voorouders. Zo gaan we een tak in de stamboom op tot we een gemeenschappelijke voorouder hebben van mensen en bv. varkens. Die tak gaan we dan afdalen tot de huidige varkens. Er is dus een keten van populaties in de geschiedenis die onze soort verbindt met eender welke andere soort. Het punt is dat de hoge morele status van A (*Homo sapiens*) vergeleken met D (een andere soort) afhangt van het toevallige feit dat tussenvormen B en C dood zijn. Het is bizar dat een criterium moreel relevant kan zijn ten gevolge van de toevallige dood van tussenvormen.

3) Ten derde is er het probleem van potentieel vage grenzen. Vele zoogdiersoorten kunnen onderling hybriden krijgen: kruisingen tussen twee soorten die vaak zelf onvruchtbaar zijn. Denk aan paard en ezel, paard en zebra, lama en kameel, leeuw en tijger, leeuw en luipaard, dolfijn en orka, schaap en geit, grizzly en ijsbeer,... Gezien de kleine genetische afstand tussen mens en chimpansee is het misschien mogelijk dat ook een mens-chimpansee hybride kan geboren worden. De vraag is dan wat de morele status van die kruising zou zijn.

En wat is de status van voorouders (de *Homo habilis*, *Australopithecus*,...)? Of van genetisch gemanipuleerde mens-dieren? Of van mens-dier chimeras? Een chimera is een individu dat bestaat uit genetisch verschillende cellen, afkomstig van zowel menselijke als dierlijke stamcellen. De lichaamscellen van chimeras kunnen variëren van 100% menselijk tot 100% dierlijk, en alles tussenin.

Deze vage grenzen zijn onvermijdelijk: wetenschappers zullen nooit een objectief criterium hebben om te bepalen of een wezen een mens is.⁴ Een ethiek die gevoelig is

⁴ Men zou kunnen tegenwerpen dat een criterium zoals voelen ook een vage grens kent tussen voelend en niet-voelend. Maar ten eerste is het een feitelijke (wetenschappelijke) kwestie of een wezen kan voelen: wetenschappers hebben reeds een aantal coherente en objectieve indicatoren (op basis van anatomie, fysiologie, gedrag en evolutionaire origine) om na te gaan of een wezen kan voelen. Ten tweede kunnen we

voor inherent arbitraire elementen, staat kwetsbaarder dan een ethisch systeem met minder arbitraire elementen. Want het risico op opportunisme wordt groter door extra arbitraire elementen in een ethiek. En een ethiek die geen raad weet met potentiële gevallen, zoals de mogelijkheid van mens-dier tussenvormen, is minder sterk dan een ethisch systeem dat bestand is tegen alle denkbare situaties.

4) De soortgrens verwijst naar genen of uiterlijk, en deze criteria zijn niet moreel relevant als het gaat over bv. huidskleur, geslacht of bepaalde lichamelijke afwijkingen. Het universalisme zegt dat we deze irrelevante moeten doortrekken naar *alle* genen en *alle* uiterlijke vormen (alle lichamelijke afwijkingen). Het is arbitrair om zonder verdere rechtvaardiging bepaalde genen of uiterlijkheden wel als criterium te nemen en andere niet. Er is bv. geen ‘belangen-gen’ dat ervoor zorgt dat enkel de wezens met dat gen complexe belangen hebben.

5) Het behoren tot de ene soort, en niet tot een andere, was niet iets wat we konden kiezen. We hebben er niet voor gekozen om als mens geboren te worden, we hebben er geen verantwoordelijkheid in, dus mogen we er ook niet voor beloond worden. Het is een schending van het verdienstenprincipe om een voorkeursbehandeling te krijgen op basis van het hebben van bepaalde genen. Als we een aparte behandeling krijgen, zou het moeten zijn omdat we bv. een inspanning hebben geleverd of een bepaalde behoefte of preferentie hebben (maar dat heeft uiteindelijk met welzijn te maken, zie volgend hoofdstuk).

De eerste vier argumenten wijzen erop dat er geen ‘essentie’ verbonden is aan een soort. *Essentialisme* wil zeggen dat er een eigenschap is die alle elementen en alleen elementen van een bepaalde verzameling (bv. een soort) hebben. Die eigenschap kan dan de verzameling definiëren. Biologisch essentialisme vinden we terug in bv. het Platonisme en bepaalde religies (bv. de idee dat alle, en alleen, mensen een ziel hebben), maar sinds Darwin is dergelijk essentialisme achterhaald. Het soortbegrip is slechts één van de vele abstracte classificaties.

Mensen zijn snel geneigd om te denken in termen van essenties (Gelman, 2003). Het essentialistisch denken wordt gebruikt in het maken van een ingroup-outgroup onderscheid (Tajfel, 1981). Psychologisch onderzoek naar racisme onthult dit fenomeen

kijken naar de neurologische basis van het bewustzijn: wetenschappers zouden er mogelijks ooit kunnen achterkomen of het ‘bewustzijnsprogramma’ aan- of afstaat. Ten derde tonen dierenwelzijnswetten aan dat we belang hechten aan gevoel bij dieren, zelfs al zou er een vage grens zijn en twijfelen we nog over bv. gevoelens bij insecten. Ten vierde tonen abortus, stamcelonderzoek en -therapie aan dat bv. een pasbevruchte embryo een lagere morele status heeft, en zouden we ook hier een eventueel vage grens van voelen (de overgang van niet-voelend embryo tot voelende foetus) aanvaarden. Waarom dan nog de soortgrens als tweede vage grens toevoegen? Een eventuele gradatie in voelen kan men nog koppelen aan een gradatie in morele status. Maar een gradatie in mens-zijn koppelen aan een gradatie in morele status wordt weer een arbitraire en vergezochte constructie. Men kan immers ook een gradatie in bv. prismaat-zijn nemen.

(Gil-White, 2001). Racisten zien rassen of etnische groepen onterecht voor essentialistische natuurlijke groepen, om zo de ingroup te kunnen afbakenen. Het is niet onwaarschijnlijk dat hetzelfde mechanisme aan het werk is bij soortisme. In discussies over dierenrechten hoort men vaak een uitspraak zoals: “De mens is de enige soort die rationeel kan denken”. Het is filosofisch gezien gevaarlijk (uiterst misleidend) om een verzameling met diverse elementen (waarvan sommigen niet aan het criterium voldoen), aan te duiden met een term in het enkelvoud. Niet alle mensen kunnen rationeel denken, een soort (als abstracte verzameling) kan al helemaal niet denken, en “dé mens” bestaat niet. Een uitspraak die verwijst naar “de mens”, blijft even waar of onwaar, even zinvol of zinloos, als we “de mens” vervangen door “de blanke” of “de primate”. De primate is immers ook een rationeel denkend wezen, in die zin dat de meeste primaten (bijna 7 miljard) een rationeel denkvermogen hebben. Maar niet alle primaten en niet alle mensen zijn rationeel denkend. Het automatisch mens-denken of soort-denken verhuult soortisme.

Het essentialisme is een intuïtieve vuistregel, een heuristiek die vaak goed werkt, maar soms niet. En het wordt cultureel beïnvloed: het toekennen van essenties aan bepaalde groepen, evenals het bepalen wie er tot de ingroup behoort, zijn cultuurgebonden (Kurzban, 2001). Onze keuze van ingroup of morele gemeenschap zit niet in onze hersenen ingebakken, ze is niet aangeboren, maar is relatief sterk cultureel-sociologisch beïnvloedbaar. Ook hier is de analogie van de Müller-Lyer-illusie treffend: die illusie ontstond ook door een intuïtieve vuistregel (3D-perspectieaanpassing) die niet altijd goed aangepast is voor tweedimensionale figuren, en niet alle culturen zijn vatbaar voor deze illusie. Net zoals er geen essentie is die alle en alleen lijnstukken met naar binnen gerichte pijlpunten kenmerken, zo is er geen essentie verbonden aan een soort. Er is geen kenmerk dat alle, en alleen, mensen bezitten.

Naast contextonafhankelijkheid, culturele afhankelijkheid en de psychologische verklaring in termen van heuristieken, was er nog een beslissend argument dat aantoonde dat de Müller-Lyer-figuur een illusie is: translatie-invariantie. De analogie van translatie-invariantie in een dierenethiek komt in het volgende deel over het welzijnsprincipe (de wijsvinger) aan bod. Uiteindelijk bekomen we zo een reeks van coherente argumenten die aantonen dat de soortgrens niet moreel relevant is en dat speciesisme een morele illusie is.

Waarschijnlijk blijven vele mensen nog vasthouden aan het intuïtieve mens-dier waardeonderscheid, zelfs nadat ze al deze argumenten hebben gehoord. Maar dergelijke hardnekkigheid hadden we ook bij de optische illusie: zelfs al meten we de lengtes, dan nog lijken de horizontale lijnstukken niet even lang te zijn telkens we een blik werpen op de Müller-Lyer-figuur. Hardnekkigheid biedt geen garantie voor waarheid of consistentie.

Het punt is dat als we blijven vasthouden aan dat mens-dier waardeonderscheid, we in conflict komen met een coherent systeem van andere morele intuïties en ethische

principes (bv. het antidiscriminatieprincipe) die sterker staan. Die argumenten verwijzen naar het arbitraire of vergezochte karakter van de soortgrens. En ons intuïtieve gevoel van morele verontwaardiging zegt dat individuen niet benadeeld mogen worden op basis van arbitraire of vergezochte criteria, want dan riskeren we te veel opportunisme. In het licht van onze geschiedenis van discriminatie (slavernij, genocides,...) is dat extreem gevaarlijk.

2. De wijsvinger: het welzijnsprincipe, onpartijdigheid en empathie

Om aan te tonen dat speciesisme een vorm van discriminatie is, is het onvoldoende om argumenten te geven voor de irrelevantie van de soortgrens. Er is een criterium nodig dat wel moreel relevant is. In dit hoofdstuk bespreek ik twee argumenten voor het criterium voelen (welzijn).⁵ Hoofdstuk drie voegt daar nog twee extra argumenten aan toe.

2.1. Onpartijdigheid

Het Rawlsiaanse gedachtenexperiment van de sluier der onwetendheid (Rawls, 1971) is een handige vuistregel om onpartijdigheid na te gaan. Stel dat je achter een sluier der onwetendheid zit, je weet dat je straks zult geboren worden in de echte wereld, maar je weet niet wie of wat je zult worden. In de originele versie van Rawls weet je dat je het leven zult leiden van een rationeel denkend wezen, maar daardoor vallen bv. mentaal gehandicapten uit de boot. Zoals Rowlands (1997) beargumenteerde, moeten we – om maximaal onpartijdig te zijn – dit gedachtenexperiment verbreden: je moet jezelf voorstellen dat je kan incarneren tot elke entiteit in het heelal. Waarom niet een planeet, elektron, inktvlek of varken in het jaar 3000? Het enige wat je achter de sluier der onwetendheid wel weet, zijn de wetten van de natuur (althans volgens de huidige wetenschappelijke kennis). Je mag nu zelf de morele wetten kiezen waarmee de

⁵ Gevoelens zijn lichamelijke of emotionele subjectieve, bewuste gewaarwordingen (qualia) die positief of negatief geëvalueerd worden, zodat ze een invloed hebben op het welzijn. Bv. het boek in mijn hand voel ik niet, tenzij ik mijn bewuste aandacht richt op mijn vingers. Dan voel ik het boek wel, maar er is niet noodzakelijk een evaluatie. Bij een pijnprikkel in mijn vinger heb ik wel een gevoel dat negatief bijdraagt aan mijn welzijn.

samenleving geordend moet worden. Iets concreter: stel dat je geboren zal worden als ofwel een mens (bv. een mentaal gehandicapte), ofwel een varken, ofwel een plant. Je weet dat ik straks één van die drie ga opeten. Je mag nu kiezen wie ik zeker niet mag opeten. Je weet ook dat volgens de huidige wetenschappelijke kennis een plant waarschijnlijk geen gevoelens heeft en dus niet vatbaar is voor welzijn. Indien je rationeel bent en bezorgd bent om je welzijn, zou je moeten beslissen dat ik noch de mens noch het varken mag eten. Want als jij de plant wordt, en als ik de plant kwets, zul jij daar niets van voelen. Je welzijn zal niet beïnvloed worden (want je zult geen welzijn hebben).

Het resultaat van dit gedachtenexperiment, is dat het criterium voelen (welzijn) moreel relevant is, omdat jij nu je welzijn waardeert. Niet-voelende entiteiten worden uitgesloten, maar je bent wel maximaal onpartijdig geweest. Dit resulteert in een consequentialistische welzijnsethiek, waarbij de gevolgen voor ieders welzijn moreel relevant zijn. Onpartijdigheid zegt dat al het welzijn van alle voelende wezens meetelt⁶. De vraag is hoeveel het meetelt. Wat is een onpartijdige verdeling van welzijn? Als je achter de sluier weet dat je het volledige leven gaat leiden van een voelend wezen, ga je belang hechten aan diens *levenswelzijn*. Dat is de waarde die je zou geven aan het leiden van het volledige leven van dat individu. Des te hoger het levenswelzijn, des te liever je dat leven zou leiden.

Als je achter de sluier een zekere risicoaversie hebt (een hogere schrik om het wezen in de laagste positie te worden), dan zou je uitkomen op een prioriteitsethiek qua verdeling van levenswelzijn, waarbij de laagste posities (met het minste levenswelzijn) een hoge prioriteit krijgen.⁷ We moeten het levenswelzijn van alle voelende wezens maximaliseren, waarbij we een sterke prioriteit geven aan het verhogen van het levenswelzijn van de wezens in de laagste posities.⁸ Met andere woorden: maximaliseer

⁶ Een voelend wezen is een wezen dat de mogelijkheid heeft ontwikkeld om te kunnen voelen, en die mogelijkheid nog niet definitief verloren heeft. Tijdens de diepe slaap is men dus nog een voelend wezen. Gewervelde dieren (en misschien ook kreeftachtigen en inktvissen) zijn volgens de wetenschap waarschijnlijk voelende wezens, omdat die een functioneel voldoende complex centraal zenuwstelsel hebben ontwikkeld.

⁷ Met maximale risico-aversie zou je absolute prioriteit geven aan het *maximaliseren* van het *minste* levenswelzijn: de maximin-regel (Rawls, 1971). Zonder risico-aversie zou je geen prioriteit geven, en zou je de totale som van ieders levenswelzijn maximaliseren: het som-utilitarisme. De prioriteitsethiek ligt tussen maximin en som-utilitarisme in. Zie ook Bruers, 2010, p291.

⁸ Wiskundig uitgedrukt zegt dit principe dat we de verwachtingswaarde van een veralgemeend (Kolmogorov) gemiddelde met concave functies van levenswelzijn van alle voelende wezens moeten maximaliseren. De maximalisatie loopt over alle beschikbare keuzes. Elke keuze die we kunnen stellen resulteert in een andere wereldgeschiedenis. In elke keuze nemen we enkel die voelende wezens in rekening die effectief bestaan of zullen bestaan in de overeenkomende wereldgeschiedenis en wiens welzijn beïnvloedbaar is door onze keuzes. De concaviteit van de concave functies in het veralgemeend gemiddelde komt overeen met de mate van risico-aversie achter de sluier der onwetendheid. Bij zeer hoge risico-aversie wordt het maximinprincipe benaderd.

het levenswelzijn van de individuen die het slechtst af zijn, tenzij dat ten koste gaat van veel meer levenswelzijn van anderen. We mogen niet te veel welzijn opofferen voor een verwaarloosbare stijging van de laagste posities, want dat zou niet efficiënt zijn. We mogen bv. niet iemand in quasi-extreme armoede drijven, om een extreme arme persoon een verwaarloosbare hoeveelheid extra welvaart te geven. We mogen ook niet iemand in armoede nog armer maken en zijn welzijn met een bepaalde hoeveelheid verlagen om zo het welzijn van welvarende personen te doen stijgen met eenzelfde hoeveelheid, want dat toont een gebrek aan prioriteit voor de personen die het ergst af zijn.

2.2. Empathie

Een tweede reden waarom het vermogen tot voelen belangrijk is, is empathie. Vanuit een deugdenethiek (en een zorgethiek) weten we dat het ontwikkelen van empathie deugdzaam is. We kunnen (op een zinnige wijze) empathie voelen met alle en alleen voelende wezens. Empathie maximaal uitbreiden kan door niet enkel mensen, maar alle voelende wezens mee te tellen.

Empathie is meestal het sterkst gericht op de ergste lijder. Diens welzijn willen we dus maximaliseren, maar niet ten koste van veel meer welzijn van anderen, want dat botst met onze behoefte aan efficiëntie. Met andere woorden, een empathische houding met een lage (maar niet afwezige) behoefte aan efficiëntie komt overeen met een onpartijdige houding (achter een sluier der onwetendheid) met een hoge (maar niet maximale) risicoaversie. We hebben dus twee verschillende gezichtspunten om naar een situatie te kijken: de rationale benadering kijkt vanuit een uitwendig, onpartijdig standpunt achter een sluier der onwetendheid waar rationele afweging belangrijk is, de emotionele benadering kijkt vanuit een inwendig, subjectief standpunt waar empathie belangrijk is.

Een deugdenethiek en een consequentialistische ethiek geven dus twee coherente argumenten voor welzijn als relevant criterium, en voor het geven van een hoge prioriteit aan het welzijn van de laagste posities. Dergelijk prioriteitsprincipe, toegepast op een dierenethiek, ligt tussen twee uitersten in: het *som-utilitarisme* van Singer (1975), waarbij de som van ieders welzijn moet worden gemaximaliseerd; en een vorm van *pijnisme* van Ryder (2001), waarbij het welzijn van de ergste lijder moet worden gemaximaliseerd.

Een hogere prioriteit wordt dus gegeven aan de laagste posities. Tot slot wordt de verwachtingswaarde genomen in geval dat de uitkomsten van de levenswelzijnswaarden onzeker zijn.

We komen zo tot ons eerste ‘materiële’ principe van gelijkwaardigheid. Volgens het formele gelijkwaardigheidsprincipe van de duim moeten we ieders welzijn een onpartijdige wijze in rekening brengen, onafhankelijk van de identiteit van individuen. Door de risico-aversie achter de sluier der onwetendheid, of door empathie, komen we tot een invulling van dit formele principe: we moeten een voorkeur geven aan het levenswelzijn van de minst bedeeden. Daaruit volgt dat, als het totale levenswelzijn van alle individuen constant is tussen verschillende situaties, de situatie met de meest gelijke verdeling van levenswelzijn de beste is. De prioriteitsethiek zit zo tussen een som-utilitarisme (dat enkel het totaal wil maximaliseren) en een egalitarisme (dat ieders levenswelzijn gelijk wil schakelen).

De sluier der onwetendheid en empathie zijn een rationeel en een emotioneel instrument om de intrinsieke waarde van een wezen (het levenswelzijn) na te gaan, door ons te verplaatsen in de positie van de ander en diens welzijn en belangen over te nemen. Het is zoals een meetlat die men verschuift van één lijnstuk naar een ander. De meetlat is een objectief (onpartijdig) instrument om een intrinsieke eigenschap van een lijnstuk (de lengte) na te gaan. En zo komen we uit bij een translatie-invariantie die we tegenkwamen bij de Müller-Lyer-illusie. Ons inlevingsvermogen (door empathie en onpartijdigheid) is een morele meetlat die we kunnen verschuiven, bv. van een mens naar een dier. Deze analogie met de Müller-Lyer-illusie maakt het plaatje nog sterker samenhangend, en is een extra reden om te geloven dat speciesisme een morele illusie is. Maar in het volgende deel over het middelprincipe geef ik nog extra argumenten waarom gevoelens moreel belangrijk zijn.

3. De middelvinger: het middelprincipe, respect en waardigheid

De consequentialistische welzijnsethiek, besproken in het vorige hoofdstuk, kan in een aantal situaties contra-intuïtieve oordelen genereren. Beschouw de volgende twee dilemma's.

1) Een vrachtwagen rijdt over een smalle weg. De remmen begeven het, en recht voor de vrachtwagen zijn vijf voetgangers. Doet de chauffeur niets of draait hij naar rechts, dan zullen de vijf voetgangers sterven. Hij kan naar links uitwijken, maar daar zal hij één fietser dodelijk treffen. Volgens het welzijnsprincipe van de wijsvinger is het beter om naar links te draaien: vijf doden is erger dan één dode.

2) In een ziekenhuis hebben vijf patiënten nieuwe organen nodig. Er is een orgaantekort, maar de chirurg kan een onbekende bezoeker opofferen en diens organen

gebruiken om de vijf patiënten te redden. Net zoals in het vrachtwagendilemma zou het wijsvingerprincipe oordelen dat gedwongen orgaantransplantatie mag. Maar onze morele intuïtie zegt van niet. Om tegemoet te komen aan deze morele intuïtie, moeten we een ander ethisch principe introduceren: de morele middelvinger.

Volgens het middelvingerprincipe heeft de onbekende bezoeker in het ziekenhuis een basisrecht om niet gebruikt te worden als louter middel voor de doelen van iemand anders. De lengte van een middelvinger is iets langer dan die van een wijsvinger, en de vingerlengte staat symbool voor het morele gewicht van een principe. Dus dit middel(vinger)principe weegt zwaarder door dan het wijsvinger-welzijnsprincipe. Bv. in het bovenstaande dilemma weegt het basisrecht minstens vijf keer zwaarder dan het recht op leven en welzijn van de vijf patiënten. De middelvinger is echter niet oneindig lang, dus het basisrecht is niet absoluut allesdominerend. Er is een zekere behoefte aan efficiëntie: gigantisch veel welzijn opofferen weegt zwaarder door dan iemand een klein beetje gebruiken als louter middel.

Het middelprincipe is gebaseerd op respect voor de waardigheid van een persoon. Maar wanneer wordt nu precies iemand gebruikt als *'louter middel'*? Er dient aan twee voorwaarden voldaan te zijn (die verwijzen naar de twee woorden 'louter' en 'middel').

1) De aanwezigheid van het lichaam van het slachtoffer is noodzakelijk om het doel te bereiken. Als het slachtoffer niet aanwezig was en het plan om een doel te bereiken zou nog steeds werken, dan wordt het slachtoffer niet gebruikt als middel. Zo kan de vrachtwagen nog steeds naar links draaien als de fietser er niet is, maar kan de chirurg niet de bezoeker opofferen als er geen bezoeker is. De bezoeker (diens lichaam) wordt gebruikt als middel (als organen), de fietser niet. Het lichaam van het slachtoffer speelt een centrale rol, want het lichaam is het enige waar men volledig eigendomsrecht over heeft.⁹ Het lichaam speelt in het gebruik als middel een centrale rol indien de lichamelijke integriteit wordt geschonden (bv. vleesconsumptie, experimenten, orgaantransplantatie, lichamelijke manipulatie), indien er een seksuele handeling is met het lichaam (bv. verkrachting), indien het lichaam een activiteit moet uitvoeren (bv. slavernij), indien het lichaam ergens moet zijn (bv. opsluiten in een kooi), indien het lichaam wordt gefotografeerd of bekeken (bv. naaktfoto, schendingen van lichamelijke privacy), indien er een prijs staat op het lichaam (bv. handel)...

⁹ We hebben de plicht om zwakkeren te helpen, maar niet als ons lichaam daarbij een centrale rol speelt. Men mag ons niet dwingen om een beetje bloed of een nier af te geven, want dat bloed en die nier zijn volledig van ons. Bloed- en nierdonaties zijn moreel goed maar niet verplicht. Men mag wel belastingen heffen (zelfs al geeft de belastingbetaler geen toestemming), want geld behoort niet tot het lichaam en is niet volledig van een persoon: het zou kunnen dat een persoon te veel geld verdiende, en als bv. een boer zijn oogst verkoopt, is de handenarbeid wel van de boer, maar de grond is (moreel gezien) niet van de boer. Iemand dwingen om arbeid te verrichten kan dan weer niet, want dan staat het lichaam wel centraal (cfr. slavernij).

2) De dader dwingt het slachtoffer bij de behandeling als middel iets te doen of ondergaan tegen de wil van het slachtoffer in. Als het slachtoffer niet graag zo behandeld wordt, wordt hij niet enkel gebruikt als middel, maar als *louter* middel. Hierbij is ook de aanwezigheid van de dader noodzakelijk. Als het slachtoffer nog steeds slachtoffer zou zijn (benadeeld zou zijn) indien er geen dader was, dan wordt het slachtoffer niet gebruikt voor het doel van de dader. Neem bv. een arme persoon die besluit brood te bakken om financieel rond te komen. Hij staat tegen zijn zin in 's ochtends vroeg op, en als ik brood ga kopen bij die bakker, gebruik ik hem wel als middel (zijn aanwezigheid is noodzakelijk), maar ik ben niet verantwoordelijk voor zijn slechte toestand. Dus ik gebruik hem niet als louter middel. Dat zou wel het geval zijn als ik hem bv. bedreig en dwing om brood te bakken; dan wordt het slavernij, want dan draag ik bij tot diens slechte toestand.

Deze twee voorwaarden genereren het middelprincipe, dat coherent is met een behoorlijk aantal morele intuïties in diverse situaties: iemand gebruiken in onvrijwillige experimenten of als menselijk schild, terreuraanslagen plegen (onschuldige burgers doden met als doel de vijand te demoraliseren), een onschuldige persoon martelen (als die persoon de locatie van een bom weet maar terroristen hem chanteren om die locatie niet te onthullen), iemand doden bij chantage (als een terrorist tegen je zegt dat jij iemand moet doden, anders gaat hij vijf anderen doden), iemand gebruiken als losgeld of lokaas (bv. bij gijzelneming), een vrouw verkrachten (en de video van de verkrachting verkopen aan duizenden pornoconsumenten), mensen verhandelen, gladiatorgevechten organiseren (ter amusement van duizenden bezoekers), exotische mensen opsluiten om ze zoals in dierentuinen tentoon te stellen (etnografische zoos), mensen doden en opeten (kannibalisme), of mensen gebruiken als slaaf. Dergelijke praktijken mogen niet, zelfs al zou het welzijn in de wereld stijgen, want telkens wordt er iemand gebruikt als louter middel.

De vraag is nu: wie krijgt dat basisrecht? Kijken we even naar de twee voorwaarden (lichamelijke aanwezigheid en onvrijwilligheid), dan moet het wezen alvast een lichaam en een wil hebben. Die wil kunnen we eng interpreteren, als de mogelijkheid om een geïnformeerde ongedwongen toestemming te kunnen geven. In dat geval krijgen enkel rationele wezens het basisrecht. Maar dan krijgen bv. mentaal gehandicapten dat recht niet.¹⁰ Een bredere invulling is de mogelijkheid om iets te kunnen willen. En alle voelende wezens die een welzijn hebben, hebben die capaciteit, want positieve en negatieve gevoelens zorgen ervoor dat een wezen iets wil (bv. pijn dat gepaard gaat met

¹⁰ Die gehandicapten hebben wel een levend lichaam, maar als we dan zouden zeggen dat alle levende wezens het basisrecht krijgen, bekomen we volgens het antidiscriminatieprincipe het probleem dat we bv. geen planten meer mogen eten. Daarom is het beter het criterium voelen te nemen.

een verlies aan welzijn, zorgt ervoor dat ik lichamelijke schade wil vermijden). Zo hebben we een extra, derde argument waarom gevoelens en welzijn moreel belangrijk zijn.

Een vierde argument zegt dat er een niet-vergezochte koppeling is tussen rechten en gevoelens: rechten beschermen belangen, en gevoelens detecteren belangen (bv. pijn zegt dat het belang van lichamelijke integriteit geschonden is, angst wijst op het belang van veiligheid). De koppeling tussen rechten en biologische soorten is daarentegen wel vergezocht.

Het middelvingerprincipe genereert een tweede materieel principe van gelijkwaardigheid: alle voelende wezens hebben een gelijke claim op het basisrecht om niet als louter middel gebruikt te worden voor de doelen van anderen.¹¹ In de dierenrechtenethiek zijn Regan (1983) en Francione (1996, 2000) voortrekkers van deze ‘deontologische’ benadering. Zo stelt Francione dat alle voelende wezens het basisrecht hebben om niet gebruikt te worden als louter bezit.

4. De ringvinger: het biodiversiteitsprincipe en het predatiedilemma

4.1. Het predatieprobleem

Een leeuw doodt in zijn leven wel honderden zebra's en andere voelende wezens. Bij predatie wordt de prooi gebruikt als louter middel, en het zou best kunnen dat predatie het welzijn sterk reduceert. Hebben we volgens het wijs- en middelvingerprincipe dan de plicht om – in de mate van het mogelijke – prooien te beschermen en predatoren uit te roeien? Zelfs vele dierenactivisten zijn daar geen voorstander van. Het predatieprobleem (Fink, 2005; Simmons, 2009) is waarschijnlijk het meest ernstige en onderschatte probleem voor zowel een consequentialistische dierenwelzijnsethiek (Singer, 1975) als een deontologische dierenrechtenethiek (Regan, 1983). Want als een leeuw nog vlees mag eten, waarom mogen wij dat dan niet?

Dierenrechtenfilosofen zoals Regan (1983) geven een onbevredigend antwoord op het predatieprobleem. Leeuwen zouden nog op zebra's mogen jagen, omdat leeuwen geen moreel besef hebben. Leeuwen zijn amoreel, niet immoreel. Maar ten eerste zouden wij

¹¹ Eventueel kunnen we aanvullen dat rationele wezens een sterkere claim op dat basisrecht hebben. Dat is niet noodzakelijk discriminatie van niet-rationele voelende wezens.

als moreel bewuste wezens nog wel een verantwoordelijkheid hebben om prooien te beschermen, net zoals we een plicht hebben om voelende wezens te beschermen tegen bv. (amorele) vallende rotsblokken. En ten tweede: wat als morgen plots leeuwen rationeel kunnen denken en ze kennis hebben van ethiek? Mogen ze dan plots niet meer jagen?¹² Of wat als op een eiland ver weg een mensenstam leeft die voor hun overleven afhankelijk is geworden van vlees en zou uitsterven als ze niet meer kan jagen? Je zou als carnivoor maar de pech hebben om een bepaalde intelligentie te hebben. Laten we daarom op zoek gaan naar een nieuw principe om predatie te rechtvaardigen. We kunnen ons hierbij laten inspireren door het carnisme, de ideologische tegenpool van veganisme.

4.2. Het 3-N-principe

Volgens sociaal psychologe Joy (2002, 2009) trachten carnisten hun vleesconsumptie te rechtvaardigen met drie argumenten: vlees is *natuurlijk*, *normaal* en *noodzakelijk*. Deze drie N-argumenten zal ik omvormen en verwerken tot een extra principe (de ringvinger) dat bepaalde vormen van predatie kan goedkeuren.

De *noodzakelijkheidsvoorwaarde* is de eenvoudigste: het gaat om een vitale behoefte voor bestaande individuen (bv. eten, medicijnen,...) of voor potentiële wezens en populaties (bv. voortplanting). *Normaal* is hier voor de eenvoud: alles wat vaak gebeurt. Het criterium *natuurlijk* definieer ik als alles wat van nature – door evolutie – ontstaan is. Evolutie is hier het gekende ‘blinde’ proces van genetische mutatie en natuurlijke selectie. Doelbewuste uitvindingen van intelligente wezens zijn dus niet natuurlijk, hoewel die intelligentie wel natuurlijk is (want door evolutie ontstaan). Natuurlijk gedrag is gedrag dat mogelijk gemaakt werd door iets (bv. een lichaamsdeel) dat via evolutie ontstond.

Het 3-N-principe (de morele ringvinger) kunnen we dan samenvatten als volgt: Als er een *(a) voldoende aantal* voelende wezens door een *(b) evolutionair proces* afhankelijk zijn geworden van een bepaald type gedrag *(c) om te overleven*, dan mogen die voelende wezens dat gedrag vertonen, zelfs al schenden ze het welzijns- en middelprincipe. Dit 3-N ringvingerprincipe zegt dus dat rechtenschendingen enkel mogen als een gedrag zowel *(a) normaal*, *(b) natuurlijk* als *(c) noodzakelijk* is. Des te normaler, natuurlijker en noodzakelijker een proces is, des te meer basisrechten mogen geschonden worden. Als slechts aan één of twee N-criteria voldaan is, dan mag het basisrecht niet geschonden worden.

¹² Mogelijks hebben dolfijnen reeds een protomoraliteit, en ze jagen op voelende vissen (Schermer, 2004, p.16).

Dit principe is coherent met een aantal morele intuïties in verschillende situaties. 1) Stel dat insecten voelende wezens zouden zijn, die we per ongeluk schaden (vertrappen) als we bewegen. Het welzijnsprincipe dreigt geschaad te worden; mogen grote wezens zoals wij dan niet meer bewegen? Jawel, want beweging is normaal, natuurlijk en noodzakelijk voor ons. Dieren doodrijden met de auto mag dan weer niet, want autorijden is niet natuurlijk; dat is een bewuste uitvinding en niet ontstaan door blinde evolutie. 2) Dieren van sommige soorten hebben een lage levensverwachting, een lage emotionele rijkdom, of hun minimale behoefte aan schaarse hulpbronnen om te overleven is erg hoog, waardoor ze competitie met andere wezens verhogen. Met andere woorden: deze wezens dragen niet goed bij tot het welzijnsprincipe. Maar ze mogen zich nog wel voortplanten, want voortplanting is als gedrag natuurlijk, noodzakelijk en normaal. 3) Men zou kunnen stellen dat verkrachting natuurlijk is (mensen hebben door evolutie een lichaamsdeel ontwikkeld waarmee ze vrouwen kunnen verkrachten, voor mannen kan verkrachting een instinct zijn) en normaal is (het gebeurt vaak in de dierenwereld, onze voorouders deden het duizenden jaren, en het is niet uitgesloten dat wij ons bestaan te danken hebben aan het feit dat een verre voorvader eens een vrouw verkrachtte). Maar het is niet noodzakelijk voor ons overleven, en het is een schending van rechten, dus mag het niet. Hetzelfde geldt voor allerlei vormen van geweld en oorlog. 4) Gedwongen orgaantransplantatie is misschien wel noodzakelijk voor de patiënten bij een orgaantekort, maar orgaantransplantatie is niet normaal (het gebeurt niet vaak) en vooral: het is niet natuurlijk (het is niet ontstaan door een natuurlijk, evolutionair proces, het is geen instinct). Dus dergelijke orgaantransplantatie die het basisrecht schendt, mag niet volgens het 3-N-principe. 5) Hetzelfde geldt voor onvrijwillige medische experimenten: dergelijk gedrag is – in tegenstelling tot predatie – niet door blinde evolutie ontstaan en bijgevolg niet natuurlijk.

Coherentie is opnieuw aanwezig, ook in dit ringvingerprincipe. Maar we kunnen met het ringvingerprincipe nog een stap verder gaan dan louter coherentie met intuïties.

4.3. De waarde van biodiversiteit

We hebben in verschillende situaties de intuïties dat anderen helpen goed is en anderen schaden verkeerd is. Er is niet enkel een coherentie tussen die intuïties in verschillende situaties, maar er bestaat ook een natuurlijke eigenschap die deze intuïties omkadert: welzijn, gedefinieerd als alle positieve gevoelens die een gevolg zijn van het goed geïnformeerd vervullen van behoeften. Kunnen we een gelijkaardige natuurlijke eigenschap vinden achter de intuïties van het 3-N-principe? Ja: biodiversiteit, gedefinieerd als alle variatie in processen (predatie,...) en entiteiten (genen, soorten,...) die een direct gevolg zijn van spontane evolutie. Zoals welzijn een complex

samenraapsel is van verschillende dingen, zo is biodiversiteit een complex samenraapsel. Zoals voelende wezens een neiging hebben om hun welzijn te verhogen door behoeftebevrediging (ondanks schaarste, trade-offs en andere beperkingen), zo hebben ecosystemen een spontane neiging om biodiversiteit te verhogen door genetische mutaties (ondanks natuurlijke selectie). Zoals voelende wezens welzijn als intrinsiek waardevolle natuurlijke eigenschap hebben, zo hebben ecosystemen biodiversiteit als waardevolle natuurlijke eigenschap.

De link tussen het 3-N-principe en biodiversiteit is als volgt in te zien. Als een proces (gedrag, eigenschap,...) natuurlijk is, draagt het per definitie bij aan de biodiversiteit. Als een proces natuurlijk en normaal is, draagt het veel bij aan de biodiversiteit. En als een proces natuurlijk, normaal en noodzakelijk is, zou de biodiversiteit drastisch dalen als dat proces er niet meer was. Als biodiversiteit intrinsieke waarde heeft¹³, is een drastisch verlies van de biodiversiteit moreel gezien heel ernstig. De waarde van een grote hoeveelheid biodiversiteit overtroeft dan het basisrecht en welzijn van individuen.¹⁴ Vandaar het belang van de 3-N-voorwaarden. Het verschil tussen predatie en gedwongen orgaantransplantatie is dus dat het eerste type van gedrag wel en het tweede niet bijdraagt aan de biodiversiteit. Bij orgaantransplantatie is een zieke patiënt niet door een blind evolutionair proces afhankelijk geworden van chirurgen en organen.

De koppeling tussen het 3-N-principe en de waarde van biodiversiteit werkt enkel als we 'de duim tegen de ringvinger plaatsen': gesteld dat we de plicht zouden hebben om deze zebra te beschermen tegen die leeuw, dan moeten we volgens de duim kunnen willen dat die plicht een universele regel wordt, dat predatie nergens en nooit mag. En dan zouden alle predatoren uitsterven (zolang ze geen voedingsalternatief hebben) en zou er een drastisch verlies aan biodiversiteit zijn. Daarom dat we in dit ene geval niet de plicht hebben om deze zebra te beschermen, zelfs al zou het beschermen van enkel deze zebra geen gevaar vormen voor de biodiversiteit. Dit is een mooi voorbeeld van de

¹³ Biodiversiteit kan alvast instrumentele waarde hebben voor voelende wezens: het kan ecosysteemdiensten leveren en bijdragen aan ecosysteemstabiliteit. Maar het is allesbehalve duidelijk of dergelijke instrumentele waarde voor het welzijn van voelende wezens voldoende is om de rechtenschendingen van predatie te rechtvaardigen. Vandaar dat de waarde van biodiversiteit misschien toch intrinsiek moet zijn en niet louter instrumenteel, om sterk genoeg te staan in het rechtvaardigen van predatie.

¹⁴ Men kan stellen dat als mensen geen vlees meer zouden eten, de veedieren dan niet meer gekweekt zouden worden, waardoor die rassen zouden uitsterven en de biodiversiteit zou dalen (Scruton, 1998). Maar 1) het bewust kweken van bepaalde veerassen draagt niet bij aan de biodiversiteit, want dat kweken is geen spontaan evolutionair proces. 2) Wilde rassen (wilde runderen,...) kunnen zich nog wel voortplanten. 3) De veeteelt zelf vormt een grote bedreiging voor de biodiversiteit (FAO, 2006). 4) Het is niet verantwoord om doelbewust ernstig lichamelijk gehandicapte wezens te kweken (de meeste veedieren hebben ernstige handicaps). En 5) dergelijk argument zouden we ook nooit inroepen ter rechtvaardiging van het kweken van mensenrassen voor slavernij.

samenwerking van de verschillende vingers om een moreel probleem zoals predatie aan te pakken.

Het probleem van het biodiversiteitsprincipe is dat veel ethici (in het bijzonder utilitaristen) zich de vraag zullen stellen waarom een niet-voelende entiteit (een ecosysteem) een intrinsieke waarde (biodiversiteit) heeft, en wat die intrinsieke waarde voor iets is. Het lijkt of we moeten kiezen tussen aan de ene kant het respecteren van bepaalde morele intuïties (bv. bij predatie en voortplanting) en aan de andere kant het vermijden van noties zoals intrinsieke waarde van niet-voelende entiteiten.

4.4. Twee andere oplossingen voor het predatieprobleem

Er zijn nog twee andere mogelijke oplossingen voor het predatieprobleem, die niet verwijzen naar een vermeende (intrinsieke) waarde van biodiversiteit.

1) Gedragsrechtvaardigheid. Een type gedrag dat rechten schendt, is toegelaten, als het gedrag noodzakelijk is, als het slachtoffer zelf schuldig is aan hetzelfde type gedrag en datzelfde type gedrag mag doen. En het type gedrag mag niet verwijzen naar een naamwoord. Als een zebra nog mag eten om te overleven, dan mag een leeuw dat ook. En dat het voedsel nu gras is of zebra's, maakt in de regel niet uit.

Dit geheel nieuwe rechtvaardigheidsprincipe is coherent met intuïties in andere situaties naast predatie. Bv.: ik offer niemand zonder toestemming op voor organen of medische experimenten, dus jij mag mij niet opofferen voor orgaantransplantatie of experimenten. Een dier van een soort die sterk bijdraagt aan het globale welzijn, mag zich nog voortplanten, dus mogen andere diersoorten zich ook voortplanten. Een insect mag nog bewegen, dus mag ik dat ook (zelfs al vertrappel ik per ongeluk een insect en zelfs al is dat insect een voelend wezen).

Een eerste probleem met dit principe is: wat als planten voelende wezens zouden zijn? De meeste planten schaden niemand bij het bewegen en eten niemand anders op, dus mogen wij geen planten meer eten. Een tweede probleem is de 'gevaarlijke wetenschapper' die tegen een biotechnoloog zegt: "Als jij experimenten mag doen, op planten, dan mag ik dat ook doen, op jou." Dit probleem kan opgelost worden door te stellen dat het argument niet mag gebruikt worden tegen rationele, morele wezens zoals biotechnologen.

2) Het bestaansprincipe. Een dader mag nog een type gedrag stellen dat rechten schendt, als dat gedrag noodzakelijk is¹⁵ en als de dader niet had bestaan indien het gedrag er nooit was geweest. Als er nooit predatie was geweest, had de leeuw nog niet

¹⁵ Deze voorwaarde is belangrijk: misschien bestaan wij slechts als gevolg van een verkrachting van een verre voorouder. We willen uitsluiten dat verkrachting zo gerechtvaardigd wordt.

eens bestaan. Als er geen orgaantransplantatie bestond, kan een patiënt die een orgaan nodig heeft nog wel bestaan.

Ook dit principe roept vragen op. Van waar komt die morele relevantie? Het lijkt vreemd dat het bestaan van het wezen zo'n belangrijke voorwaarde is. En zouden in sommige gevallen (onvrijwillige) experimenten wel toegelaten zijn? Indien mijn grootvader in leven kon blijven dankzij het resultaat van experimenten, heb ik mijn bestaan te danken aan dergelijk gedrag, en mag ik zelf (onvrijwillige) experimenten doen? Dit lijkt me een ernstig bezwaar te zijn voor het bestaansprincipe.

Bovenstaande drie oplossingen van het predatieprobleem hebben enkele problemen. Het 3-N-biodiversiteitsprincipe lijkt misschien toch de sterkste kandidaat voor de ringvinger te zijn. Maar we kunnen ook het 3-N-principe en het gedragsrechtvaardigheidsprincipe combineren tot een derde materieel gelijkwaardigheidsprincipe: alle wezens hebben een gelijk recht op gedrag dat natuurlijk, normaal en noodzakelijk is (gedrag dat sterk bijdraagt aan de biodiversiteit).

5. De pink: het partijdigheidsprincipe en getolereerde keuzegelijkwaardigheid

Een onpartijdige ethiek zoals we die zagen in de vorige drie morele vingers, is soms erg veeleisend. Daarom introduceer ik een vijfde basisprincipe, dat een zekere vorm van partijdigheid toelaat. Metaforisch gesproken: de pink kan een klein beetje afwijken van de andere vingers, maar de pink is de kleinste vinger, dus te veel partijdigheid mag niet. Hoeveel partijdigheid we dan wel mogen hebben, kunnen we afleiden door de duim tegen de pink te plaatsen: we mogen partijdig zijn tot op zeker niveau, op voorwaarde dat we gelijkaardige niveaus van partijdigheid tolereren bij iedereen in alle gelijkaardige situaties.

Laten we, om in de sfeer van predatie te blijven, dit principe toepassen op het volgende dilemma: een groep hongerige leeuwen komt aangerend; je hebt de tijd om ofwel je eigen kind ofwel twee onbekende Afrikaanse kinderen te redden. Het wijsvingerprincipe wijst naar het redden van de twee onbekende kinderen, maar toch zouden we tolereren dat je je eigen kind redt. Dergelijke partijdigheid willen we toelaten. En er is meer: als jij jouw kind redt in plaats van de andere kinderen, mag ik je daarvoor niet veroordelen, want dan beschouw ik je als louter middel. Je aanwezigheid is immers noodzakelijk om de twee andere kinderen te redden, en je zou dan iets doen tegen je wil. Dus het middelvingerprincipe zorgt ervoor dat ik je keuze moet tolereren,

waardoor de pink een beetje ruimte krijgt om af te wijken van het wijsvingerprincipe. Er is dus een zekere samenhang tussen de morele vingers.

Dit pink-partijdigheidsprincipe kan ingeroepen worden om een aspect van het predatieprobleem op te lossen: zowat alle mensen zouden wel een kind en niet een zebra redden van de leeuwen¹⁶. Maar dat hoeft geen speciesisme te zijn. Als blanke je eigen kind redden, en niet een Afrikaans, getuigt ook niet noodzakelijk van racisme. Je zou wel racist zijn als je mijn keuze om het Afrikaanse kind te redden niet zou tolereren (als je bv. zou zeggen dat jouw kind blank is en het daarom meer rechten heeft om geholpen te worden). Hieruit besluiten dat je een ander kind mag opofferen om met diens hart het jouwe van een hartkwaal te genezen, is echter een stap te ver. Besluiten dat we dieren mogen eten, omdat we geneigd zijn in noodsituaties eerst mensen te redden, is net zo onterecht.

Het partijdigheidsprincipe beantwoordt de vraag geformuleerd in de titel van een boek van Gary Francione: “Your child or the dog?” (Francione, 2000) Dit verwijst naar een brandend-huis-dilemma dat vegetariërs vaak te horen krijgen. Stel dat een huis in brand staat, en je hebt slechts tijd om ofwel je eigen kind, ofwel de hond, te redden. Wanneer men dit dilemma stelt, is de achterliggende idee vaak dat een dierenactivist consequent onpartijdig moet zijn. Hij zou een muntstuk kunnen opgooien om te bepalen wie hij zal redden: één kans op twee dat het de hond wordt, ten koste van het eigen kind. De dierenactivist die verkiest zijn eigen kind te redden – wat erg waarschijnlijk is – wordt beschuldigd van inconsequent gedrag. Het pink-partijdigheidsprincipe verworpt deze beschuldiging.

Net als bij het dilemma waarbij je moet kiezen tussen jou kind en een onbekend kind, is er in dit brandend-huis-dilemma een *gevoelsmatige ongelijkwaardigheid* als je de voorkeur geeft aan je eigen kind. Maar het partijdigheidsprincipe genereert wel een subtiele vijfde vorm van gelijkwaardigheid (naast de vier vormen besproken bij de vorige vingerprincipes): *getolereerde keuzegelijkwaardigheid*. Dit is gebaseerd op het tolereren van de keuze van andere hulpverleners. Stel dat jij in het brandend huis jouw kind redt en niet het mijne. Als wij elkaar als gelijkwaardig beschouwen, dan ‘erven’ onze kinderen een vorm van gelijkwaardigheid. Ik beweer namelijk niet dat je mijn kind moet redden omdat mijn kind meer morele waarde en rechten heeft. Gevoelsmatig heeft mijn kind wel meer waarde voor mij, maar ik zou jouw keuze om jouw kind te redden respecteren. Een gevoelsmatige ongelijkwaardigheid kan dus nog wel gepaard gaan met een getolereerde keuzegelijkwaardigheid.

De regel van getolereerde keuzegelijkwaardigheid zegt dus dat men bij het helpen van anderen wel een zekere prioriteit mag geven aan de wezens met wie men een

¹⁶ Uit persoonlijke contacten met talrijke dierenactivisten bleek dat zelfs *antispeciesisten* hiertoe neigen.

sterkere emotionele betrokkenheid voelt, op voorwaarde dat men de keuze van andere hulpverleners respecteert die prioriteit geven aan hun dierbaren. Getolereerde keuzegelijkwaardigheid komt tegemoet aan de kritiek van zorgethici op de kille onpartijdigheid van het welzijnsprincipe. Net zoals een antiracist tolerant moet zijn tegenover de keuze voor het redden van het Afrikaanse kind, moet een antispeciesist de keuze tolereren om een hond en niet een mens te redden. Zoiets gebeurt misschien wel vaker dan we denken: iemand die een hond als huisdier heeft, zorgt meer voor die hond dan voor de kinderen van iemand anders. Men geeft bv. meer geld uit aan eten voor de eigen hond dan aan medicijnen voor een kind in Afrika.

Belangrijk bij getolereerde keuzegelijkwaardigheid is dat de zorgverlening of hulpactie gemotiveerd wordt door een gevoel van empathie en betrokkenheid en niet door een vooroordeel.

Conclusie: veganisme als morele plicht?

Toegepast op onze consumptie van dierlijke producten, vertelt de morele hand een interessant verhaal.

De wijsvinger wijst in de huidige veeteelt en visvangst een ernstige schending van het welzijn aan. Achter de sluier der onwetendheid zou je nooit kiezen voor een dergelijk systeem als je weet dat je geboren kan worden als veedier. En het dierenleed botst ook met onze waarden van mededogen en empathie.

Volgens de middelvinger is zowat elke vorm van veeteelt en visvangst onverantwoord (niet enkel de intensieve, maar ook de meer 'romantische' vormen), omdat we zowat altijd het lichaam van voelende wezens gebruiken op een manier waarbij ze dat niet willen. Als men – met enige fantasie – mensen zoals mentaal gehandicapten zou kweken en gebruiken zoals we dieren gebruiken in de meest diervriendelijke veeteelt, zouden we het nog steeds verafschuwen.

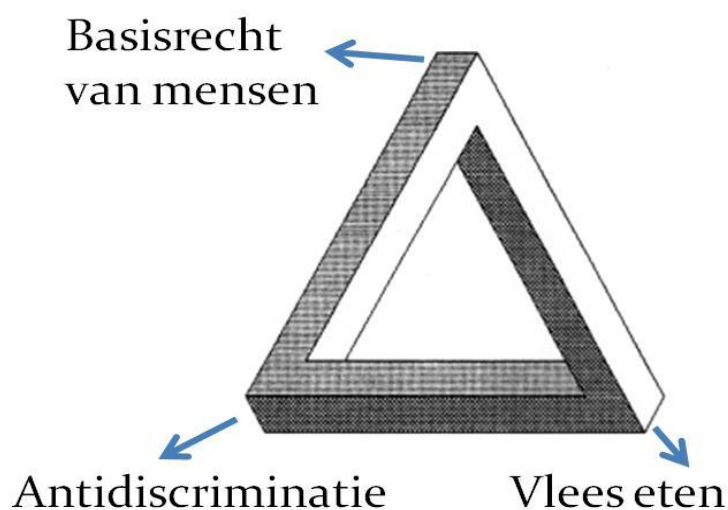
Het ringvinger-biodiversiteitsprincipe invoeren om de rechtenschendingen en het verlies aan welzijn te rechtvaardigen, zal niet lukken, want de consumptie van dierlijke producten is niet noodzakelijk voor ons. Vlees eten mag dan wel, net zoals verkrachting, in zekere zin normaal en natuurlijk zijn, maar we hebben geen dierlijke producten, en ook geen verkrachting, nodig om gezond te leven. De grootste voedingsdeskundigenorganisatie in de wereld, de Academy of Nutrition and Dietetics, stelt letterlijk dat een goed gepland plantaardig voedingspatroon gezond is voor alle

mensen, inclusief zwangere vrouwen, jonge kinderen, topsporters¹⁷,... (ADA, 2009). De biodiversiteit gaat niet dalen als we veganist worden. Integendeel, de veeteelt vormt zelfs een ernstige bedreiging van de biodiversiteit (FAO, 2006).

En volgens het pinkprincipe zouden we nooit de vorm van partijdigheid die nodig is om veeteelt te kunnen verantwoorden, tolereren. Getolereerde partijdigheid inroepen lukt dus ook niet.

Veganisme is dus ethisch consistent. Tot slot impliceert het duimprincipe dat we het goede voorbeeld moeten geven: zelfs al haalt het niets uit als enkel jij en ik besluiten om veganistisch te leven (de markt gaat daar niets van merken), dan nog blijft het onze morele plicht en moeten we doen wat iedereen zou moeten doen.

Als afsluiter wil ik een andere optische illusie voorleggen: de onmogelijke drie-balk, symbool voor het interne conflict in onze ethiek. De drie hoekpunten stellen ethische principes en waarden voor die gebaseerd zijn op morele intuïties. Het eerste principe zegt dat alle voelende mensen rechten hebben, zoals het basisrecht van de middelvinger. Het tweede hoekpunt zegt dat we nooit mogen discrimineren, en het derde zegt dat we dieren mogen gebruiken om op te eten (voor de smaak, want voor de gezondheid is het niet nodig).



Figuur 16: de onmogelijke drie-balk van de vleeseter

We zijn ons niet bewust van de inconsistentie, omdat we telkens gefocust zijn op één of twee hoekpunten, terwijl we niet uitzoomen om te kijken of het globale plaatje klopt. Als we het geheel overschouwen, stellen we vast dat die drie principes onderling niet verzoenbaar zijn: als je niet mag discrimineren (tussen bv. mens-dier) en als je wel vlees mag eten, dan is bv. kannibalisme toegelaten. Eén van de principes moet sneuvelen.

¹⁷ Zie bv. veganistische kampioenen zoals Carl Lewis, Patrik Baboumian, Scott Jurek en Mac Danzig.

Ofwel mogen we mensen zonder hun vrijwillige toestemming gebruiken als louter middel (bv. gedwongen mensproeven, kannibalisme, mensenhandel, verkrachting, slavernij), ofwel mogen we discrimineren (is ook racisme tolereerbaar), ofwel mogen we geen vlees en andere dierlijke producten meer eten. Als we kijken naar de geschiedenis en de horror van slavernij en racisme, is het zeer waarschijnlijk dat we de eerste twee principes absoluut niet willen afschaffen. Het derde principe, het tolereren van vleesconsumptie, staat het zwakst, en zal omgedraaid moeten worden: we mogen geen vleesconsumptie tolereren. Het derde hoekpunt in de figuur moeten we dan ook omkeren.

Appendix: a review and systematization of the trolley problem

Abstract¹

The trolley problem, first described by Thomson (1976) and Foot (1978), is one of the most famous and influential thought experiments in deontological ethics. The general story is that a runaway trolley is threatening the lives of five people. Doing nothing will result in the death of those persons, but acting in order to save those persons would unavoidably result in the death of another, sixth person. It appears that, depending on the situation, we have different moral judgments about the permissibility of action. We will review and systematize all the proposals in the literature of the past 35 years that have attempted to grasp our moral intuitions in a simple deontological principle. In particular, seventeen proposals will be classified: six algorithmic, seven psychological, and four other invalid accounts. This review and classification sheds light on some subtle differences and clarify a few issues.

Introduction

The trolley problem consists of a series of moral dilemmas involving a runaway trolley threatening the lives of a certain number of people.² The basic structure of all the

¹ This appendix is based on Bruers & Braeckman (2013).

² For a highly readable overview and historic background of 'trolleyology', see Edmonds (2013).

dilemmas is the same: if you do not act, five people will die; if you act, one other person will be killed and the five will be saved. Research into the way people deal ethically with the trolley dilemmas has shown that most people's intuitions do not correspond either with pure (extreme) deontology or with utilitarianism (Greene et al., 2001; Waldmann & Dieterich, 2007; Hauser et al., 2008). By 'pure' deontology we mean here, for simplicity's sake, that people should comply with the following rule: never act if the act results in harming people who were not threatened if you had not acted. By 'pure' utilitarianism we mean that people should comply with the rule: always choose the action that maximizes the number of lives saved (i.e., least total harm).

Different trolley dilemmas have the same consequential structure but yet, confronted with those dilemmas, people hardly ever say that one should never act, or that one should always act. When presented with different dilemmas, most people say that we must act in one trolley situation, but in another dilemma we are not allowed to act; it is as if people make inconsistent choices. Only pure utilitarian consequentialism states that we should always act in all the trolley dilemmas. So most people's moral intuitions deviate from these consequentialist ethics, and therefore the trolley problem is an interesting thought experiment for studying deontological ethics. The basic question is the following: What is the morally relevant difference between Dilemma A and Dilemma B, such that it is morally allowable to act in A, but not to act in B? Also, in this article we state that a consistent moral solution of the trolley problem should contain a clear description of a rule or principle that best fits, justifies, and explains the diversity of people's moral intuitions in the diverse cases. In other words, the best solution to the trolley problem is a clear algorithm to decide whether one should act or not, and the answers that this algorithm generates should be in line with intuitions.

Many people have tried to solve the abovementioned basic question. In this article, we present an overview of the many proposals that ethicists have come up with during the last 35 years, and we discuss their differences, mutual relations, strengths, and weaknesses. In addition to covering the most relevant versions of trolley dilemmas and the solutions proposed in the literature, we also present some new hypothetical solutions. But probably the most important contribution of this article to the existing literature is a systematic classification of all those solutions. And looking at new trolley dilemmas, we clarify the differences between the proposed solutions (principles).

There are several reasons why this new classification is important. First, it certainly helps to avoid confusion between different solutions (we will mention some confusions in the literature). Second, the classification of the different principles gives us insights into which of those principles and underlying moral intuitions could be something like 'moral illusions' (e.g., Unger, 1996). Third, people adhering to deontological ethics might be able to see which proposed principle they would most prefer; that is, which of the proposals is most compatible with their own moral intuitions. Fourth, our findings will have implications for further empirical studies in moral psychology (e.g., Greene, 2002;

Cushman et al. 2006; Mikhail, 2007; Greene, 2008). This systematization opens up some new questions. Do people prefer one of the proposed solutions? How many people would agree with which solution? Would they change their judgments in some dilemmas in order to make them fit with their preferred solution? And if there are different proposed solutions related to different (psychological or algorithmic) mechanisms, does that mean that there would be more ‘moral modules’ in our brains (e.g., the brain research on trolley dilemmas done by Greene et al., 2001)?

We start this review with a number of trolley dilemmas that cover all the important issues and elements that are discussed in the literature. Then, we select six possible solutions to the trolley problem that are described in the literature (the sixth is in fact a new solution), from which we will suggest that these can be grouped in pairs, so that there are in fact only three groups of principles with more or less strong support in the literature. These accounts have an algorithmic character, with a clearer and more objective decision procedure than the other, psychological accounts.

Of course, other hypothetical principles are possible, but they have few or no supporters or they remain dubious and are still debated. So after describing the three groups of ‘algorithmic’ accounts, we give an overview of seven other proposals that do not distinguish permissibility from non-permissibility so clearly. Some of these other proposals might be relevant, as they are more ‘psychological’ in nature and psychology strongly influences our moral judgments. However, these psychological explanations are not always clear or do not always make a distinction *between* the different ‘agent-neutral’ trolley dilemmas: they make distinctions *within* one dilemma. An agent-neutral trolley dilemma is a description of a situation that excludes agent-related information.³ The inclusion of agent-related information such as a person’s position (e.g., distance from the victim) or mental state (e.g., knowledge, risk attitude, intention to harm) give rise to further distinctions within the same agent-neutral trolley dilemma.

After the six algorithmic and seven psychological accounts, we finally briefly highlight four other proposals encountered in the literature that do not solve the trolley problem, because they result in pure deontology. In summary, seventeen proposals in the literature are classified as follows: six solutions that make objective (algorithmic) distinctions between different dilemmas; seven solutions that make distinctions within one and the same dilemma, depending on some psychological state; and four invalid proposals that always result in pure deontology in all dilemmas.

So let’s start the trolley’s engine.

³ The only (trivial) agent related information in all situations is that if the agent acts, it is supposed that s/he acts with the intention or plan to save the people on the main track. I.e. malicious intentions (e.g. to kill a hated person) are excluded.

The trolley dilemmas

In this section we will briefly present and systematize the most commonly discussed versions of the trolley dilemma (for further details, see Thomson, 1985; Kamm, 1989, 1998; Otsuka, 2008; Fischer & Ravizza, 1992b).

Dilemma 1: The switch. A trolley is moving towards five people on the main track. You are standing at a switch. If you turn the switch, the trolley will be diverted to a side track, but there is one person on this side track. Turning the switch will result in that person's death, and the five people on the main track will be saved. Should you turn the switch? Most people (roughly 90% according to Hauser et al., 2008) say you are allowed to do so.

Dilemma 2: The bridge. A fat man is standing on a bridge above the track. You can save the five people on the track below by pushing the fat man from the bridge in front of the trolley, so that the trolley will be stopped by his heavy weight. The fat man will die, and the five people will be saved. Only a few people (roughly 10% according to Hauser et al., 2008) say that we are allowed to push the fat man. Most people either refuse to push the fat man or condemn pushing the fat man. According to Waldmann and Dieterich (2007), people are more tolerant of pushing someone onto the tracks, but in their study, the dilemmas did not involve close up and personal contact with the victim who is pushed. Their dilemmas looked more like Dilemma 5 below, where the victim is in a truck, so you do not have to touch the victim personally.

Dilemma 3: The loop. As in the first dilemma, you are standing at a switch. But this time the side track turns back onto the main track. If there is no one on the side track, the trolley will still move onto the main track and will kill the five people. But on the side track is a fat man. So if you turn the switch, the fat man will block the trolley. The fat man dies, the five people will be saved. In a recent survey, Hauser et al. (2008) found that roughly half of the respondents said that turning the switch is permitted. However, according to Waldmann and Dieterich (2007), people are more tolerant towards turning the switch. But they constructed the dilemma in a different way, where the person – the fat man – on the side track is sitting on a bus. So on the side track the bus will block the trolley, not the fat man. The fat man in the bus will die in the accident.

Dilemmas 1 to 3 share a rather similar basic structure, as can be seen in **Figure 17** below. In the first picture, the people on the main track are standing between two forks and the trolley is situated before the first fork. This is equivalent to Dilemma 1 (the switch). In the second picture, the five people are behind the second fork, and you have hesitated so long that the trolley already passed the first fork. You could still save the people on the main track, because the side track is at a height, so you could easily push the fat man from the side track onto the main track. In the third picture, the people on the main track are behind the second fork, as in Dilemma 3 (Loop), and you still have

time to turn the switch. The difference is in the positions of the trolley and the people on the main track.

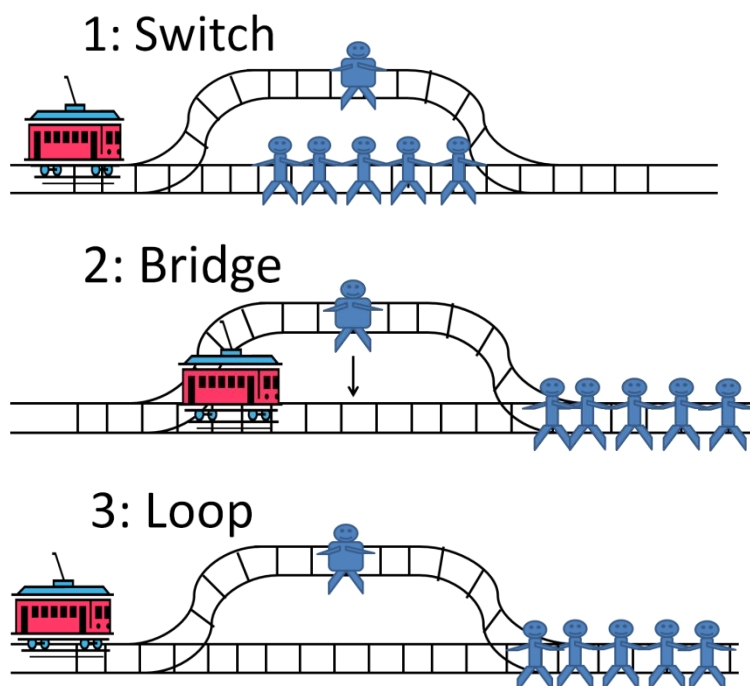


Figure 17. Basic structures of Dilemma 1 (Switch); Dilemma 2 (Bridge); and Dilemma 3 (Loop).

Dilemma 4: The loop with a stone. The situation resembles the one in Dilemma 3. But this time a heavy stone is located behind the man on the side track. The man on the side track is not heavy enough to stop the trolley. The stone will block the trolley if you turn the switch. But the man in front of the stone will die. In a survey performed by Hauser et al. (2008), it was found that nearly three quarters of the respondents say that turning the switch is allowed, and this is a statistically significant difference from Dilemma 3.

Dilemma 5: The truck. You can block the trolley by pushing a heavy truck onto the rails. In this truck there is one passenger, who will get killed (this dilemma was studied by Waldmann & Dieterich, 2007).

Dilemma 6: The rockslide. You can turn a switch, redirecting the trolley onto a side track. On this side track there is a big rock. When the trolley hits the rock, the rock slides towards a bystander and kills him.

Dilemma 7: The platform. Five people are on a moving platform on the rails. If you do nothing, the trolley will crush the platform and kill the five. But you can move the platform away from the rails in order to save the five. But this move will push another person (who is standing next to the platform) on to an electric cable. This person will consequently die by electrocution. This dilemma is similar to the 'Lazy Susan' case in Kamm (1989).

The next dilemmas is new in the literature, and will be used to point out differences between some accounts.

Dilemma 8: The loop with an avalanche. This dilemma is similar to the loop dilemma, but the person on the side track is controlling a safety barrier against avalanches. The person on the side track is not heavy enough to stop the trolley, so one needs to create an avalanche. But only after the person dies is it possible to initiate an avalanche that is not blocked by the barrier; so only after the victim dies and is no longer controlling the barrier, can the trolley be blocked and the five saved.⁴

Six algorithmic accounts

Ethicists have looked for morally relevant differences between the above dilemmas. They want to find a moral rule that generates answers that are consistent with the answers (intuitive moral judgments) of the majority of people. As we have seen, the majority of people are very clear about the first two dilemmas, the switch and the bridge. Therefore, we think that a good solution should consist of a most precise formulation as possible of a moral criterion to distinguish the dilemmas, such that action is at least permissible in the switch dilemma but not in the bridge dilemma. Concerning the other dilemmas, there is less consensus about people's moral intuitions, so the solutions might differ in these cases. We prefer a moral rule that works like a kind of algorithm; that is, a clear procedure applicable to all dilemmas, which provides an unambiguous answer as to whether action is allowed or not, and without reference to fuzzy or ambiguous concepts.

There are five algorithmic accounts proposed in the literature. We will introduce a sixth. They can be grouped together in pairs; hence, we have structured them into three groups of accounts. The first proposal in each group is a rather vague account, vulnerable to misinterpretations or borderline cases. The second account in each group is more precise, leaving less room for interpretation. In other words, the second accounts correspond to more accurate interpretations of the first ones in each group. We will apply these algorithmic accounts to the above dilemmas and show that these three explanations are different from each other.

⁴ We note that Lippert-Rasmussen (1996) gave another dilemma that has a resemblance to Dilemma 8, according to the accounts mentioned in the next section.

Group A: the ‘mere means’ accounts

A1: Use as merely a means to an end. The Kantian inspired right not to be treated solely as a means is based on the unalienable dignity of persons. This right trumps the right to life of other persons: it is never allowable to kill and use a person as merely a means – even if this means that by this act the lives of others could be spared. It is considered disrespectful to treat someone merely as means. This mere means account is mentioned in Thomson (1985), and as we will see in a later section, it is related to an interpretation of the doctrine of double effect (Quinn, 1989b).

Looking at the dilemmas, we can say that the fat men on the bridge (Dilemma 2) and on the loop track (dilemma Loop) are used as ‘trolley blockers’ or ‘human shields.’ So only in these dilemmas is action not allowed (they violate the dignity of the victim). In the other dilemmas action is permitted.

This account can sometimes be a bit vague, as it is not always easy to understand what use as merely a means really is. Sometimes one might have to use an element of fantasy to refer to an analogous means or instrument. We believe the following account is equivalent to this mere means account, but it provides a more algorithmic way (a clear test) to decide whether the Kantian right is violated or not.

A2: The counterfactual account about the required presence of the victim (mentioned in Thomson, 1985; Parfit, 2011). If the presence of the potential victim is (causally) required in order to save the five people on the main track (i.e., if it would be impossible to save the five people without the victim’s body), then you should not act. Here we can easily decide whether the Kantian right is violated, by asking ourselves what would happen if the one person in the trolley dilemmas was not present. If nobody is on the bridge, it is impossible to push someone in front of the trolley to stop it. But in considering Dilemma 1 (Switch), saving the five would still be possible if the one person on the side track had not been present. In Dilemma 4 (Loop and stone) one can still turn the switch and let the trolley be stopped by the stone. In Dilemma 5 (Truck) one can still move the truck, even if there is nobody in it. The platform (Dilemma 7) can still be shifted when there is no one standing next to it. In Dilemma 8 (Loop with avalanche) one could still start an avalanche when the victim on the side track is not present.

According to Waldmann and Dieterich (2007), using a person as a means is not a criterion that people use, because a lot of people say that we are allowed to turn the switch in Dilemma 3 (Loop). However, in their study, the person on the side track was sitting in a bus, and people might think it was the bus that is blocking the trolley. So the bus is used as the means, not the person’s body. If the person was not sitting on the bus, saving the five by turning the switch still works. Furthermore, Hauser et al. (2008) noted a slight difference in people’s responses between Dilemmas 3 (Loop) and 4 (Loop and stone). This difference can only be explained by the mere means accounts.

The mere means account also has another property: if the victim's body needs to be present (if there is no other heavy object that could replace the person as a trolley blocker), it also implies that there is logically no possibility of saving the one person after the five people on the main track are saved. After turning the switch in Dilemma 4 (Loop with stone), you could still try to save the person on the side track. But in Dilemma 3 (Loop), saving this person is impossible: even if you manage to run to the person and pull him away from the tracks, you cannot do this without endangering the five people again. This property might point at an evolutionary explanation of the moral intuition: in rescuing members from your group, it is advantageous to choose the option that allows you to try to save all of them. Saving everyone is not logically impossible in Dilemma 1 (Switch) and Dilemma 4 (Loop with stone). You first save the five people and then run to the side track. If you do not run fast enough, you still have saved the five, and if you can run fast enough, you can save everyone. The latter is logically impossible in Dilemmas 2 and 3 (Bridge and Loop).

Compared with the next four accounts, the mere means accounts are the most reliable: they are more accurate, have less boundary cases and generate less judgments that are strongly counter-intuitive in some dilemmas (the only counter-intuitive judgment occurs in the loop dilemma).

Group B: the 'same threat' accounts

The two same threat accounts that we are about to discuss have something in common; they both claim that it is only permissible to act if two conditions are satisfied: (1) no new threat is introduced, but a pre-existing threat is redirected or redistributed from the larger to the smaller group (this is also referred to as the Permissible Diversion Hypothesis in Postow, 1989); and (2) another condition is satisfied – about this latter condition, we will discuss two candidates (related to rights or interventions), but we expect that they are equivalent.

The first condition is not satisfied in Dilemma 6 (Rockslide) and Dilemma 7 (Platform), because electrocution and rockslides are new threats. So in these dilemmas, action is already not permitted. However, a problem of this condition is its lack of clarity: it is not always clear when a new threat is introduced. There are some borderline cases of redirected threats that more resemble new threats (e.g., situations where trolleys change after being sent to a side track). These borderline cases might undermine the objective, algorithmic nature of the same threat account. Leaving this issue aside, let's look at the second condition. There are two versions of the second condition, leading to two accounts.

B1: Violation of rights. Thomson (1976, 1985) had the idea that action in the case of Dilemma 2 (Bridge) is not allowable, because pushing the fat man is an infringement of an important right. On the other hand, in Dilemma 1 (Switch) turning the switch does not violate a similar right of the person on the side track because you do not do anything to him. In particular, you do not push him, so his right not to be pushed is not violated. In other words in Dilemma 2 (Bridge), you do something to a person (which is a violation of rights), whereas in the switch you act on the threat (which is not a rights violation). Also in all the loop dilemmas (3, 4, 8) nobody is pushed, so no right not to be pushed is violated. Turning the switch is itself morally neutral and not a violation of rights. Therefore, Thomson claimed that it is permitted to turn the switch in the loop dilemmas.

Thomson's idea has been criticized as being too vague and for contradicting moral intuitions (Kamm, 1989; Postow, 1989). What rights are we talking about – the right not to be killed by trolleys or the right not to be pushed? The following describes another candidate condition, which in fact might be equivalent to what Thomson had in mind, but stated a bit more clearly.

B2: Sending a victim to the trolley. Some ethicists claim that there is a morally relevant difference between throwing a bomb at a person and throwing a person at a bomb; or in trolley language, sending a trolley to a person versus sending a person to a trolley. This is referred to as intervention myopia (Waldmann & Dieterich, 2007) and focuses at the locus of intervention: do you in the first instance intervene in the path of the aggressor (the trolley, the bomb) or the path of the victim? This criterion has some supporters (Boorse, 1994; Harris, 2000; Waldmann & Dieterich, 2007) and some critics (Fischer, 1992; Fischer & Ravizza, 1994). Montmarquet (1982) also offered a same threat principle, but this approach was criticized by Kamm (1989).

Only in the bridge and truck dilemmas does one send the victim (the fat man or the passenger) to the trolley. In these cases, action is not allowed, even if the threat is the same. We also note that in Dilemma 7 (Platform), the victim is sent to a new threat (electrocution), so both conditions of the same threat account are not satisfied. Yet, we expect that most people's intuitions would allow action in this dilemma. This gives a strong counter example to the same threat account.

Furthermore, there are some boundary cases between sending victims and threats to each other. As an example, consider a loop dilemma whereby turning the switch also shifts a platform, positioning the victim exactly on the side track to block the trolley. Also, according to Unger (1996, p. 101), the difference between sending the victim to the trolley versus sending the trolley to the victim is an illusion, based on what he called

‘protophysics.’⁵ In his book, Unger (1996) also gives other similar irrelevant protophysical differences that influence our moral judgments. For example: in some dilemmas it is worse to save some people and harm someone else by increasing the speed of a trolley than by decreasing it.

Group C: the ‘causal chain’ accounts

The next two principles look at the causal chain that is the result of action or inaction. We note that these principles should be taken with a grain of salt, because a clear and consistent interpretation of them might just be impossible if we think about them more critically. Nevertheless, we present them here.

C1: Principle of (Im)Permissible Harm (PI/PH). We cite Kamm (1989, in Darwall, 2003, p. 167), who introduced this hypothesis: “It is permissible to cause harm to some in the course of achieving the greater good of saving a greater number of others from comparable harm, if events which produce the greater good are not more intimately causally related to the production of harm than they are to the production of the greater good.” This is a complicated formulation that needs more explication, so let us look at the dilemmas to see what is meant by “intimately causally related.”

Looking at the switch dilemma, the action is turning the switch, and this action has two consequences that appear at the same instant in the causal chain: the five are saved, and the one is threatened by the trolley. The production of the harm (the threat to the one on the side track) is causally related to the action of turning the switch, and also the saving of the five is causally related to the turning of the switch. Both are in this dilemma equally intimately causally related to the turning of the switch, because both are the direct consequences of turning the switch. The condition of the PI/PH is satisfied, so it is allowable to turn the switch.

In the bridge dilemma, however, the action is pushing the fat man. As a first consequence, the fat man is threatened; a second consequence is that the fat man blocks the trolley and saves the five. But looking at the causal chain, we see that the ‘causal distance’ between the action (the pushing) and the harm (or the threat) to the fat man is smaller than the causal distance between the action and the saving of the five. The

⁵ The loop dilemma is often used by some philosophers (e.g., Singer, 2005; Scanlon, 2008) to demonstrate the invalidity of the deontological mere means account, the abovementioned mere means principle, by claiming that a lot of people have the intuition that it is permissible to act in the loop dilemma, even when the victim is used as a trolley blocker. However, if this protophysical explanation is correct, the judgment in the loop dilemma (the permissibility to act), might be a moral illusion. We will demonstrate in more detail in another study.

saving of the five happens further up in the causal chain. Therefore, PI/PH says that action is not permitted.

As it is not always clear how to calculate intimate causal relatedness, there are some borderline cases. In Dilemmas 3 (Loop), the situation is similar to Dilemma 1 (Switch), according to Kamm (1989).⁶ However, we might disagree with this, as can be seen in Situations 2 and 3 (Figure 17). In both the bridge and the loop dilemmas, the fat man is simply placed in the path of the trolley, either by changing the path of the trolley (Situation 3, Loop) or changing the position of the fat man (Situation 2, Bridge). Causally speaking, both are equivalent. So we should be a bit skeptical about this account.

The sixth possible explanation is not mentioned in the literature (as far as we are aware), and perhaps it is identical to an interpretation of Kamm's PI/PH hypothesis above. It is also vulnerable to borderline cases, such as the loop dilemma.

C2: The increased threat account. This rule says: you are allowed to act if the victim of your action dies (is harmed) after the others are saved. In the causal chain, we can represent it as follows. Each person has a value: 1 equals *alive and saved*, 0 means *dead (or harmed)*, X means *actually threatened (but still alive)*, and Y means *potentially threatened* (meaning it is possible to turn a threat towards that person). The six persons in the dilemmas all have a value at each step; so we can represent the starting situation as X,X,X,X,X,Y, that is, five persons are actually threatened and one person is potentially threatened (we are able to act so that his position would become a threatened one). Turning the switch in Dilemma 1 changes the situation to Y,Y,Y,Y,Y,X, which means that one person is really threatened and the five people are potentially threatened: we can turn the switch back to change the situation back to the initial situation. So the five people are not yet absolutely sure about their survival. At a particular point, when the trolley passes the bifurcation in the track, the five people are actually saved and the situation turns into 1,1,1,1,1,X. And after a few moments, the person on the side track dies, resulting in 1,1,1,1,1,0. So the causal chain in the switch dilemma looks like:

$$X,X,X,X,X,Y \rightarrow Y,Y,Y,Y,Y,X \rightarrow 1,1,1,1,1,X \rightarrow 1,1,1,1,1,0.$$

Action is allowable if the causal chain looks like that above. However, in Dilemma 2 (Bridge), matters are more complicated: it all depends on whether the fat man is heavy enough to block the trolley (i.e. whether the five are definitely saved once you pushed the fat man). It might be the case that the trolley is too fast and is able to kill all six people, because all six people are placed in the trajectory of the trolley. In other words,

⁶ In her later work, Kamm (2007) introduced new refinements (e.g. causal versus non-causal flip sides, directly versus indirectly causing a lesser evil, producing versus sustaining a greater good and substituting versus subordinating persons). But this was criticized by Norcross (2008) as being heavily ad hoc and unclear.

it is not clear that the five people are absolutely saved *already at the moment when one pushes the fat man*. The causal chain now can look like:

$$X, X, X, X, X, Y \rightarrow X, X, X, X, X, X \rightarrow 1, 1, 1, 1, 1, 0.$$

The possibility of the X,X,X,X,X,X situation (everyone is in danger), distinguishes the bridge from the switch dilemma. The number of threatened people is increased. In Dilemma 8 (Loop with avalanche) we clearly see a moment where everyone is in danger: when turning the switch and the trolley passes the bifurcation, the one person on the side track is threatened, but the five other people are also still threatened, because the one person is not heavy enough to block the trolley. Only after the one person is killed does it become possible to relieve the threat to the five, by initiating an avalanche that blocks the trolley. The causal chain in Dilemma 8 looks like:

$$X, X, X, X, X, Y \rightarrow X, X, X, X, X, X \rightarrow X, X, X, X, X, 0 \rightarrow 1, 1, 1, 1, 1, 0$$

The causal chain account is only able to distinguish Dilemma 1 (Switch) from Dilemma 2 (Bridge) if the causal chain can have a point where everyone is in danger, for example if we suppose that the train might kill all six people in Bridge. The existence of this increased threat situation in the causal chain disallows action. But then we have to suppose a similar possibility in Dilemma 3 (Loop), which disallows action. However, Dilemma 4 (Loop and stone) becomes creates a boundary case: if the stone was a real mountain, the possibility of situation X,X,X,X,X,X is as unlikely as it is in Dilemma 1 (Switch). So the permissibility depends on whether the stone is really heavy enough to block the trolley. But the same can be said about the fat man, who might be heavy enough to block the trolley. If we know the fat man is heavy enough, there is no distinction between Switch and Bridge: after pushing the fat man, the path of the trolley changes (it stops), just as the path of the trolley changes in the switch dilemma. Pushing the fat man and turning the switch automatically guarantee the immediate safety of the five people. We can introduce a distinction by claiming that the agent cannot be sure whether the fat man is heavy enough, but this turns the algorithmic account into a psychological account (see the section in risk aversion below).

The increased threat account reveals a kind of ‘causal myopia’ (similar to the term ‘intervention myopia’ related to the same threat account; Waldmann and Dieterich, 2007). If in the series of consequences of your action you do not threaten someone before or at the moment when others are really saved, then you are allowed to act. It is as if you were blind to the further consequences in the causal chain.

Table 1 presents the results of the trolley dilemmas according to the above three principles: the mere means, same threat, and causal chain accounts. A plus means that the action is allowed, a minus means that it is not. As discussed, the causal chain accounts have some question marks.

Dilemma	Mere means account	Same threat account	Causal chain account
1. Switch	+	+	+
2. Bridge	–	–	–(?)
3. Loop	–	+	–(?)
4. Loop and stone	+	+	–(?)
5. Truck	+	–	–(?)
6. Rockslide	+	–	+
7. Platform	+	–	+
8. Loop and avalanche	+	+	–

Table 1: answers to the trolley dilemmas, according to the three accounts.

Note that the switch and bridge cases get all plus and minus signs respectively, so for these all three accounts can be considered as a solution to the trolley problem. But the answers differ when looking at other dilemmas.

Seven psychological accounts

The accounts that we presented above are all objective, in the sense that they did not refer to mental states, but to events, counterfactual requirements, number of threats, points of intervention, directions or causal consequences. In this section, we give an overview of some other proposals discussed in the literature. These proposals often involve some psychological influences, such as intentions, risk aversion, personal versus non-personal conflict, and so on. These psychological accounts have some flaws: sometimes they are not able to derive a clear judgment in a certain dilemma (especially the loop dilemma generates problems of interpretation), or they do not always make a clear distinction *between* dilemmas such as the switch and the bridge. They make distinctions even *within* one dilemma. So depending on the situation (related to the psychological states) it might be possible that it is not permitted to act in the switch dilemma, or that it is permitted to act in the bridge dilemma.

1. The Doctrine of Double Effect (DDE). This doctrine is mentioned in quite a few discussions about the trolley problem (Boyle, 1980; Davis, 1984; Fischer & Ravizza, 1992a; Reibetanz, 1998; McIntyre, 2001; Shaw, 2006; Edmonds, 2013). The doctrine says that there is a moral difference between the intentional harm as a means and the foreseen

harm as a side-effect (Quinn, 1989b). It has been criticized by, for example, McIntyre (2001).

The DDE is an agent-centered, psychological account, as it makes a difference between what the agent intends or foresees.⁷ We could try to interpret the DDE in a more agent-neutral way; that is, without too much reference to the mental states of agents. Reinterpreting the DDE as an agent-neutral principle moves it close to the mere means account discussed above, because the DDE refers to 'harm as a means.' However, we have to be careful not to confuse the use of a person's body as a means versus the use of, for instance, a switch as a means or a plan as a means.

The difference between the DDE and the mere means accounts can be most clearly seen in Dilemma 8 (Loop and avalanche). The person on the side track is not used as a means, because the presence of his body is not necessary to save the five (on the contrary, his presence has prevented the initiation of the necessary avalanche). But the agent intends the killing (removal) of the person on the side track, because this removal is necessary in order to initiate the avalanche. The DDE says that action is not allowed, because it involves an intentional harm.

Hence, the DDE is not simply equivalent to the mere means account, a fact that might result in misinterpretations in the literature. For example Costa (1986), in his application of the DDE to the trolley dilemma, combined (or confused?) the mere means account with a version of the causal chain account. And to make it even more extraordinary (or confusing), in a later article Costa (1987) also included a version of Thomson's 'same threat' principle in the description of the DDE, as if the DDE is a confusing mixture of all three groups of accounts discussed in the previous section.

The major problem with the DDE is the loop dilemma: is the death of the person on the side track intended or merely foreseen? When Kamm (2000) tried to apply the DDE to the loop trolley dilemma, she promoted a new doctrine of triple effect (DTE).

2. The Doctrine of Triple Effect (DTE). Following Kamm's doctrine, turning the switch in the loop case is permissible according to triple effect. That is because apart from intentional harm (doing something *in order to* bring about an evil) and merely foreseeing a side effect (doing something *in spite of* bringing about an evil), Kamm claims that there is a third option, in which one does something *because* it brings about an evil (which should be distinguished from 'in order to bring about an evil'). This DTE approach was further defended by Shaw (2006) but criticized by Harris (2000) and more

⁷ Of course, the objective accounts also include a trivial mental state of the agent: if the agent acts, s/he is supposed to have an intention or plan to save the people on the main track. However, the DDE refers to a non-trivial mental state: the intention to harm (distinguished from foreseeing the harm).

recently by Otsuka (2008) and Liao (2009) using the loop dilemma: triple effect does not solve the loop case either. Liao argued that the because of/in order to distinction does not apply to the loop case, and furthermore questions whether this distinction has a normative significance.

Otsuka (2008) gives an example of a trolley dilemma where this triple effect becomes clearer: suppose you are at a switch, and on the side track there is one person in front of six other people. If you turn the switch, the five on the main track are saved, the first person on the side track will block the trolley, and the six people behind him are saved. Here, we can say that we would turn the switch, not *in order to* kill the one on the side track, but rather *because* he will be killed and stop the trolley. Nothing new is added however, we think, because action in this dilemma is also allowed according to our abovementioned three accounts.⁸

3. *Feelings of the victim.* Thomson (1993) invited us to focus on what the potential victim would feel about what the agent does. If you were thrown from a bridge you might feel differently about the agent, than if a trolley were directed towards you. And it is this difference that plays a role. However, this claim also involves some complex knowledge of psychology, this time not of the agent, but of the victim. It does not yet solve the trolley problem, because one can imagine switch and bridge situations where the victim feels the same.

4. *Projective grouping.* Peter Unger speculated about another psychological mechanism behind our moral judgments: projective grouping and projective separating (Unger, 1996, p. 97). “[When certain people are in a situation that is taken to be their problem], we tend to think it is badly wrong to spare them the serious losses that might stem from their problem by imposing serious loss on other people, who don’t have that problem.” In the first trolley dilemma (Switch), all six people on the tracks are considered to be in a similar position in that they have something in common: they are all on a track and could be run over by a trolley. So the five on the main track and the one on the side track are grouped together as having the same problem, and the one on the side track can therefore be considered as ‘fair game’ to be sacrificed. However, in the bridge dilemma, the fat man is in a different position: he is not on a track, but on a bridge. So the fat man is psychologically separated from the five people on the track, which makes us decide not to sacrifice the fat man. A lot of people, when responding to

⁸ For some further subtlety, however, we can say that the one person on the side track is a means to save the six behind him, but he is not used as means to save the five. If the person was not present, the plan to turn the switch and save the five would still work (but six other people would be threatened).

the trolley dilemmas, give spontaneous answers that reflect this projective separation (people say something like, “But the fat man had nothing to do with it, he was just passing by”). Also Hanna (1992) proposed a Principle of Moral Inertia, which is basically the same as the projective separating. A distinction is made between participants (such as the person on the side track in the switch dilemma) who are part of an ongoing causal process, and bystanders (such as the fat man on the bridge) who are not part of the ongoing process. But this explanation is not fully satisfactory, however, because as Unger himself argued, it can be twisted. And it is at the least very vague: there are no clear criteria to separate people into groups. There is no consensus about what the relevant differences should be. Knowing whether someone is a participant or a bystander is not straightforward. And what about Situation 2 in **Figure 17**, where the fat man was on a side track on a bridge?

5. Epistemic accounts: risk aversion. Risk aversion is a psychological attitude that might give an interesting explanation for the moral intuitions in the trolley problems. Can we know whether our plan to save the five would really work? If the fat man is not heavy enough and the trolley were to keep on moving, then all six will die, which is an even worse outcome. There is the risk of a worse outcome. If the trolley could have stopped in time, even without the fat man blocking it, then the fat man would have died unnecessarily. In the switch dilemma, however, we can be pretty sure that the five are saved and nobody dies in vain.

According to this epistemic account, action would be impermissible if there is a possibility that the rescue plan will fail and all six people will die. In particular, action might not be allowed in Dilemmas 2, 3, and 5 (Bridge, Loop and Truck).

The problem with this hypothesis is that certainty is a matter of degree. Take Dilemma 4 (Loop and stone): What if the stone was really heavy so that you could be sure that it would stop the train? Surely a mountain of stones would be convincing. And even in the switch dilemma, suppose that the side track bends behind a hill. You cannot be sure that there are no people on the side track behind the hill. Perhaps there are ten people on the side track, but you cannot see them.

So the epistemic account in fact makes distinctions even within one dilemma, instead of between dilemmas. Nevertheless, there might be some interesting truth in this approach. It is related to the amount of risk aversion that the agent has. Suppose in the bridge dilemma there is a 10% probability that the plan of pushing the fat man fails and all six people die instead of one, an 80% probability that the plan will work and one person will die instead of five, and a 10% probability that the trolley could have stopped anyway without the fat man, so that one person dies instead of nobody. A person with a high level of risk aversion would choose not to act. A person with maximum risk aversion would never act, even if the probability of failure were 0.0001%. In this context, we note that most people have a high but not maximum level of risk aversion.

Going back to the switch dilemma, risk aversion would imply not turning the switch if there is a possibility that there are ten more people down the side track. But be aware that the same could apply to the main track: it might *equally* be possible that there are ten people behind the five, and you did not see them. Not turning the switch would result in fifteen deaths. Notice the word 'equally'. There is a kind of symmetry in the switch dilemma; whereas in most bridge dilemma situations that we imagine we do not see such a symmetry, and risk aversion has a stronger influence in those dilemmas.

6. *Epistemic accounts: uncertainty aversion.* Next to risk aversion there is uncertainty (or ambiguity) aversion, whereby the probabilities of success are not even known. The probability that the plan involving pushing the fat man will work is not 10%: it is usually not known. So we have to choose between two games of chance. Suppose that you are one of the six people in the bridge trolley dilemma, but you do not know which one. If the fat man is not pushed, you know that the trolley will continue moving and kill five people. So you have a survival probability of one sixth, because you have a one sixth probability of being the fat man who survives. This is the first game of chance. In the second game, the fat man is pushed, and there is still a possibility that the trolley continues on and kills one or more of the five people on the track. Perhaps all might die. Which game of chance would you prefer to play? The situation is very similar to Ellsberg's paradox (Ellsberg, 1961). Suppose we have an urn and you know three things: it contains six balls, has six (or fewer) different colors, and there is one green ball. The choice is between two games of chance. In the first, you win when you draw the green ball. Your probability of winning is one sixth. In the second, you win when you draw a blue ball. Your probability of winning is now unknown (somewhere between zero and five sixths), because you do not know how many blue balls there are. Some (or most) people prefer the first game, because they have uncertainty aversion. The similarity with the trolley game is obvious.

7. *Personal versus impersonal dilemmas.* Greene (2008), finally, points – using psychological and brain research – to an important aspect in the trolley dilemmas: the distinction between personal versus impersonal dilemmas, related to the relative position of the agent towards the victim. Pushing the fat man is an action, which is close up and personal, whereas turning the switch is a more detached action. This is certainly something that influences people's choices, but it is not sufficient to solve the trolley problem, because it is easy to invent scenarios such as the bridge dilemma to make the action more detached (e.g., you are standing far away from the bridge and the fat man, but you can push a button, overturning the bridge). So this criterion would also make a distinction within the bridge dilemma. When most people imagine the bridge dilemma as a close up and personal situation, some emotion reaction in their brains will be triggered and tell them not to push the fat man.

Interestingly, in their research, Greene et al. (2001) classified personal dilemmas using some criteria, one of them reads: “where this harm is not the result of deflecting an existing threat onto a different party” (Greene, 2002, p168). This refers to the same threat account.

Four invalid accounts

In this section, we summarize some proposals that in fact would all result in ‘pure’ deontology, so they do not solve the trolley problem.

1. The Doctrine of Doing versus Allowing (DDA). This principle of DDA is that there is a moral difference between killing and letting die. Quinn (1989a) for example, referred to the DDA to distinguish between trolley dilemmas. But as Fischer and Ravizza (1992a) argued, matters get very complicated in applying the DDA to trolley dilemmas, because one needs to include unsatisfactory references to concepts such as ‘transfer of intentions,’ ‘causal isolation,’ and so on.

In line with the DDA, Foot (1978) made a distinction between positive duties (aid) versus negative duties (non-interference), and applied this to the trolley dilemmas. But this approach was criticized by Thomson (2008), who showed that Foot’s idea basically results in pure deontology, whereby no action is permitted in all the trolley dilemmas.

Interestingly, Thomson (2008) also ends up with pure deontology. However, we think that she is mistaken at some point. Thomson (2008, p.365) used a wrong argument (wrong analogy) to demonstrate that action is never allowed. Let us digress on this a little, because it is a recent discussion. Thomson starts with the ‘three options’ dilemma: you are at a switch – if you do nothing the trolley will kill five people on the main track. You can also turn the switch to the right hand track, where one person will be killed, or turn it to the left hand track, where you will be killed. The argument goes that nobody is willing to sacrifice himself/herself (apart from real altruists or depressed people), and it is really unfair, Thomson claims, to turn the switch to the one victim on the right hand track. To show that this is unfair, Thomson uses another example: You are asked to give money to a charity, in order to save people. You are able to send your own money, but you instead feel like stealing the money of someone else and sending that money to the charity. We claim that the analogy does not apply, because in the charity dilemma, you are *using* something of someone else. It is comparable to the ‘transplant dilemma’ (Thomson, 1985), whereby a surgeon can save five patients by sacrificing an innocent person and use that person’s organs for transplantation. The transplant dilemma is

similar to the bridge trolley dilemma, whereby you also use something of the victim, namely his body, without his consent. So Thomson's analogy can be used to argue that pushing the fat man in the bridge dilemma is not allowed. But from the charity dilemma analogy, it does not yet follow that turning the switch is not allowed. FitzPatrick (2009) and Shaver (2011) also commented on Thomson's new turn towards pure deontology (Thomson, 2008).

2. Illegitimate plans. Russell (1979) referred to 'illegitimate plans' to argue against the permission of, for example, pushing the fat man from the bridge. But this idea was criticized by Kamm (1989), and is in fact equal to pure deontology, as all the actions are shown to be illegitimate plans.

3. Threatened persons. Montmarquet (1982) stated that only when a person was not threatened is action impermissible. But he claimed that the person on the side track is already threatened. This claim, however, is false, as was argued by Gorr (1990). Montmarquet's approach would result in pure deontology, just like the DDA (Gorr himself, by the way, refers to a 'same threat' account).

4. Rare situations. Gert (1993) claimed that – in contrast with impermissible actions – the permissible actions occur in very rare situations. Sitting in a truck next to a railway (Dilemma 5) or standing on a bridge above a railway (Dilemma 2) are more typical situations than standing on the railway with no escape possible, standing next to a moving platform, or standing in the pathway of a rock, and so on. The problem with this approach is that it is difficult to quantify the rarity of a situation and derive from this the permissibility of actions.

Conclusion and further research

The basic question we now have to ask is whether our pattern of answers (following our moral intuitions) is given by one or more of the three accounts presented above: the mere means, same threat, or causal chain accounts. If not, there must be other principles, or we must have moral illusions (comparable to optical illusions and cognitive biases). We know that there are some very peculiar examples of irrationalities in people's answers to the trolley dilemmas. Unger (1996) demonstrated that people's responses to the trolley dilemmas are often more (inconsistent) psychology than ethics, by pointing out that judgments about the permissibility of an option (e.g. the choice to push the fat man) depend on the availability of other options that people consider as

being irrelevant (see also Norcross, 2008). There is the well known effect of wording and framing (Petrinovich & O'Neill, 1996; Sinnott-Armstrong, 2008; Lanteri et al., 2008; Ray & Holyoak, 2010). Especially the order in which different trolley dilemmas are presented, has some influence (Petrinovich & O'Neill, 1996; Liao et al. 2011; Schwitzgebel & Cushman, 2012; Di Nucci, 2012). And also induced feelings of disgust (Schnall et al., 2008) and happiness (Valdesolo & DeSteno, 2006) can influence moral intuitions in the trolley dilemmas. The turn that the trolley problem has made towards empirical studies in moral psychology (Greene, 2008; Cushman et al. 2006; Mikhail, 2007) is very fruitful, especially in discovering moral illusions. Still, the abovementioned studies in experimental philosophy do not indicate that the gap between the two paradigmatic cases Switch and Bridge can be closed.

Given the classification of different accounts above, we can now ask the following (empirical) questions: How many people can agree with one or more of the three principles? Which account will have the most followers? What happens if respondents learn about these accounts? Do people feel satisfied with these accounts, and would they pick a preferred one? Will this influence their judgments in some dilemmas, and how? If, for example, a person is permissive towards the action in the loop dilemma, but learned that the mere means account is perfectly compatible with all of his/her intuitions in all dilemmas, except for the loop dilemma, would that change the judgment in the loop dilemma? Can such reflections easily override intuitions? Would that eventually influence the intuition in that dilemma? Can those accounts be used in the method of reflective equilibrium (Rawls, 1971)? These can be questions for future research.

Bibliography

- ADA (2009). Position of the American Dietetic Association: Vegetarian Diets, *Journal of the American Dietetic Association*, 109(7):1266-82.
- Adams, C. (1995). *The Sexual Politics of Meat*, Continuum, New York.
- Adams, C. (1995b). *Neither Man nor Beast: Feminism and the Defense of Animals*, Continuum, New York.
- Adams C. & Donovan J. (1996). *Beyond Animal Rights: A Feminist Caring Ethic for the Treatment of Animals*, Continuum Intl Pub Group.
- Ahluwalia, A. (1978). An intra-cultural investigation of susceptibility to 'perspective' and 'non-perspective' spatial illusions. *Br. J. Psychol.*, 69(2): 233-241.
- Allais, M., (1953). Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école Américaine, *Econometrica*, 21: 503-546.
- Allen, M.W., Wilon M., Ng, S.H. & Dunne M. (2000). Values and beliefs of vegetarians and omnivores. *Journal of Social Psychology*, 140 (4): 405-422.
- Allen, M. W., Gupta R. & Monnier A. (2008). The Interactive Effect of Cultural Symbols and Human Values on Taste Evaluation. *Journal of Consumer Research*, 35: 294-308.
- Arneson, R.J. (1989). Equality and Equal Opportunity for Welfare. *Philosophical Studies*, 56: 77-93.
- Arneson, R.J. (2008). Equality of Opportunity. *Stanford Encyclopedia of Philosophy*. Fall 2008 Edition.
- Arrhenius, G. (2000). *Future Generations: A Challenge for Moral Theory*, PhD dissertation, Uppsala University.
- Arrhenius, G., Ryberg, J. & Tännsjö, T. (2010). The Repugnant Conclusion, *The Stanford Encyclopedia of Philosophy* (Fall 2010 Edition), Edward N. Zalta (ed.).
- Arrow, K.J. (1963). *Social Choice and Individual Values*, 2nd ed., New Haven and London: Yale University Press.
- Arrow, K.J. (1965). *The Theory of Risk Aversion*, in *Aspects of the Theory of Risk Bearing*, by Y.J. Saatio, Helsinki.
- Atkinson, A.B. (1970). On the Measurement of Economic Inequality. *Journal of Economic Theory*, 2 (3):244-263.
- Beauchamp, T.L. & Childress, J.F. (2001). *Principles of Biomedical Ethics*, 5th ed. New York: Oxford University Press.
- Bekoff, M. (2007). *The Emotional Lives of Animals: A Leading Scientist Explores Animal Joy, Sorrow, and Empathy and Why They Matter*. New World Library.
- Bekoff, M. & Pierce, J. (2009). *Wild Justice: The Moral Lives of Animals*. University of Chicago Press
- Benson, J. (2001). *Environmental Ethics: An Introduction with Readings*. London: Routledge.
- Bernstein, M. (2004). Neo-speciesism. *Journal of Social Philosophy*, 35 (3):380-390.

- Blackorby, C., Bossert W., & Donaldson, D. (2002). Population Principles with Number-Dependent Critical Levels, *Journal of Public Economic Theory*, 4:347–68.
- Blackorby, C., Bossert W., & Donaldson, D. (2003). The Axiomatic Approach to Population Ethics. *Politics Philosophy Economics*, 2(3): 342–381.
- Blackorby, C., Bossert W., & Donaldson, D. (2005). *Population Issues in Social Choice Theory, Welfare Economics, and Ethics*. Cambridge University Press.
- Bloom, P. (2010). Why We Like What We Like. *Observer*, 23 (8):3.
- Bognar, G. & Kerstein, S.J. (2010). Saving Lives and Respecting Others. *Journal of Ethics & Social Philosophy*, 5(2):1–20.
- Boorse, C. (1994). Ducking Trolleys, *Journal of Social Philosophy*, 25(3):146–152.
- Bostrom, N. (2006). Quantity of experience: brain-duplication and degrees of consciousness. *Mind Mach* 16:185–200.
- Bostrom, N. & Yudkowski, E. (2011). The Ethics of Artificial Intelligence. In Ramsey W. and Frankish, K. (eds.) *Cambridge Handbook of Artificial Intelligence*, Cambridge University Press.
- Boyle, J.M. (1980). Toward Understanding the Principle of Double Effect. *Ethics* 90(4): 527–538.
- Broome, J. (1991). *Weighing Goods*, Oxford; Basil Blackwell.
- Broome, J. (2004). *Weighing Lives*. Oxford: Oxford.
- Brophy, M. (2009). *Moral Intuitions in Reflective Equilibrium: Applying Scientific Methodology to Ethics*. Dissertation, University of Minnesota.
- Brown, C. (2007). Prioritarianism for Variable Populations. *Philosophical Studies*, 134:325–361.
- Bruers, S. (2013a). Speciesism as a Moral Heuristic. *Philosophia*, 41(2): 489–501.
- Bruers, S. (2013b). Dieren, dilemma's en discriminatie: naar een samenhangende ethiek van dierengelijkheid. *Ethiek & Maatschappij*, 14 (4):33–59.
- Bruers, S. en Braeckman, J. (2013) A review and systematization of the trolley problem. *Philosophia*, DOI: 10.1007/s11406-013-9507-5.
- Byrne, A. (2010). Inverted Qualia, *The Stanford Encyclopedia of Philosophy*, Spring 2010 Edition.
- Carruthers, P. (1992). *The Animals Issue: Moral Theory in Practice*, Cambridge University Press, Cambridge.
- Carter, A. (1999). Moral Theory and Global Population, *Proceedings of the Aristotelian Society*, 99: 289–313.
- Changizi M.A., Hsieh A., Nijhawan R., Kanai R. & Shimojo S. (2008). Perceiving the Present and a Systematization of Illusions. *Cognitive Science*, 32(3):459–503.
- Chappell, T. (2011). On the Very Idea of Criteria for Personhood. *Southern Journal of Philosophy*, 49 (1):1–27.
- Chomsky, N. (1986). *Knowledge of language: Its nature, origin, and use*. New York, NY: Praeger.
- Clark, S. (1977). *The Moral Status of Animals*, OUP.
- Cohen, C. (1997). Do Animals Have Rights? *Ethics and Behavior*, Harvard University, 7(2):91–102.
- Cohen, C & Regan, T. (2001). *The Animal Rights Debate*. Lanham, MD: Rowman & Littlefield.
- Cohen, G. A. (1989). On the Currency of Egalitarian Justice. *Ethics*, 99:906–944.
- Cohnitz, D., Häggqvist, S. (2009). The Role of Intuitions in Philosophy. *Studia Philosophica Estonica* 2:1–14.
- Cosmides, L., Tooby, J. & Kurzban, R. (2003). Perceptions of Race. *Trends in Cognitive Sciences*, 7(4), 173–179.
- Costa, M.J. (1986). The Trolley Problem Revisited. *The Southern Journal of Philosophy*, 24(4):437–149.
- Costa, M.J. (1987). Another Trip on the Trolley. *The Southern Journal of Philosophy*, 25(4):461–166.
- Cover T.M. & Thomas, J. A. (1991) *Elements of Information Theory*. Wiley-Interscience.
- Cowen, T. (2003). Policing Nature. *Environmental Ethics*, 25 (2):169–182.
- Crisp, R. (2008). Well-being. *Stanford Encyclopedia of Philosophy*. Edward N. Zalta (ed.).

- Cushman, F., Young, L., & Hauser, M. (2006) The Role of Conscious Reasoning and Intuition in Moral Judgment: Testing Three Principles of Harm. *Psychological Science*, 17(12):1082–1089.
- Daniels, N. (1979). Wide Reflective Equilibrium and Theory Acceptance in Ethics. *Journal of Philosophy*, 76(5):256–82.
- Darwall, S. (ed.) (2003). *Deontology*. Oxford: Blackwell Publishing.
- Davis, N. (1984). The Doctrine of Double Effect: Problems of Interpretation. *Pacific Philosophical Quarterly* 65: 107-123.
- Davis, S. (2003). Least Harm. *Journal of Agricultural and Environmental Ethics*, 16(4).
- Dawkins, R. (2004). *The Ancestor's Tale*, Boston: Houghton Mifflin.
- Deutsch, D. (1992). Paradoxes of Musical Pitch. *Scientific American*, 267 (2): 88–95.
- Devine, P.G. (2001). Implicit Prejudice and Stereotyping: How Automatic Are They? *Journal of Personality and Social Psychology* 81(5): 757–759.
- Dick, J. (1975). How to Justify a Distribution of Earnings. *Philosophy and Public Affairs*, 4: 248-72.
- Di Nucci, E. (2012). Self-Sacrifice and the Trolley Problem. *Philosophical Psychology*, forthcoming.
- Dombrowski, D. (1997). *Babies and Beasts. The Argument from Marginal Cases*. University of Illinois Press, Chicago.
- Donaldson S. And Kymlicka W. (2011). *Zoopolis. A Political Theory of Animal Rights*. Oxford University Press, USA
- Dorsey, D. (2009). Headaches, Lives, and Value. *Utilita*, 21:54-6.
- Dougherty, T. (2013). Aggregation, Beneficence and Chance. *Journal of Ethics & Social Philosophy*, 7(2):1-19.
- Dworkin, R. (1981). What is Equality? Part 1: Equality of Resources. *Philosophy and Public Affairs*, 10: 185-246.
- Ebert, R. & Machan, T. (2012). Innocent Threats and the Moral Problem of Carnivorous Animals. *Journal of Applied Philosophy*, Vol. 29, No. 2, pp:146-159.
- Edmonds, D. (2013). *Would You Kill the Fat Man? The Trolley Problem and What Your Answer Tells Us about Right and Wrong*. Princeton University Press.
- EFSA (2009). *General approach to fish welfare and to the concept of sentience in fish*. Scientific Opinion of the Panel on Animal Health and Welfare, EU Food Safety Authority.
- Ehrlich, P. and Holdren, J. (1971). Impact of Population Growth, *Science*, 171:1212-1217.
- Elster, J. and Roemer, J. (1991). *Interpersonal Comparisons of Well-Being*. Cambridge University Press, Cambridge.
- Ellsberg, D. (1961). Risk, Ambiguity, and the Savage Axioms. *Quarterly Journal of Economics*, 75 (4): 643–669.
- Epstein, L.G. (1999). A Definition of Uncertainty Aversion. *The Review of Economic Studies*, 66 (3): 579.
- Everett, J. (2001). Environmental ethics, animal welfarism and the problem of predation; A Bambi lovers respect for nature. *Ethics & the environment*, 6(1):42-67.
- Fairlie, S. (2007). Can Britain Feed Itself? *The Land*, 4.
- FAO (2006), *Livestock's Long Shadow*. United Nations Food and Agriculture Organisation, Rome.
- Fink, C.K. (2005). The Predation Problem. *Between the species*, Issue V:1-16.
- Finnis, J. (1995). A Philosophical Case Against Euthanasia. In: Keown J. (ed.) *Euthanasia Examined*. Cambridge University Press.
- Fischer, J.M. (1992). Thoughts on the Trolley Problem. In: *Ethics: Problems and Principles*, Fischer, J. M. & Ravizza, M., New York: Holt, Rinehart & Winston.
- Fischer, J.M. & Ravizza, M. (1992a). Quinn on Doing and Allowing, *The Philosophical Review*, Vol. 101, No. 2.
- Fischer, J. M. & Ravizza, M. (1992b). *Ethics: Problems and Principles*. New York: Holt, Rinehart & Winston.

- Fischer, J. M. & Ravizza, M. (1994). Ducking harm and sacrificing others, *Journal of Social Philosophy*, 25:3.
- FitzPatrick, W.J. (2009). Thomson's Turnabout on the Trolley. *Analysis* 69(4):636–643.
- Fleurbaey, M. (2009). *Assessing risky social situations*. LSE Choice Group working paper series, vol. 5, no. 9.
- Fleurbaey, M., Tungodden, B. & Vallentyne, P. (2009). On the possibility of nonaggregative priority for the worst off, *Social Philosophy and Policy*, 26:258–285.
- Fleurbaey, M. & Zuber, S. (2012). Inequality aversion and separability in social risk evaluation. *Economic Theory*. DOI 10.1007/s00199-012-0730-2.
- Foot, P. (1978). The Problem of Abortion and the Doctrine of Double Effect. In: *Virtues and Vices*. Oxford: Basil Blackwell, pp.19–32 (originally appeared in the *Oxford Review* 5, 1967.)
- Fox, M.A. (1999). *Deep Vegetarianism*. Temple University Press.
- FRA (2010). *The Right to Political Participation of Persons with Mental Health Problems and Persons with Intellectual Disabilities*. European Union Agency for Fundamental Rights, Vienna, Austria.
- Francione, G. (2000), *Introduction to Animal Rights: Your Child or the Dog?* Philadelphia: Temple University Press.
- Gajdos, T. & Kandil, F. (2008). The Ignorant Observer. *Social Choice and Welfare*, 31(2):193–232.
- Gelman, S. (2003). *The Essential Child: Origins of Essentialism in Everyday Thought*. New York: Oxford University Press.
- Gert, B. (1993). Transplants and Trolleys. *Philosophy and Phenomenological Research*, LIII(1):173–179.
- GFN, (2010). National Footprint Accounts, Edition 2010, data year 2007. Global Footprint Network, Oakland, USA.
- Gilboa, I. & Schmeidler, D. (1989). Maxmin Expected Utility with Non-Unique Prior. *Journal of Mathematical Economics*, 18(2):141–153.
- Gilligan, C. (1982). *In A Different Voice*. Cambridge MA: Harvard University Press.
- Gil-White, F.J. (2001). Are ethnic groups biological “species” to the human brain? Essentialism in our cognition of some social categories. *Current Anthropology*, 42(4): 515–555.
- Gödel, K. (1931). Some basic theorems on the foundations of mathematics and their implications. In S. Feferman, ed., 1995. *Kurt Gödel Collected works, Vol. III*. Oxford University Press.
- Godlovitch, R., Godlovitch S. & Harris J. (eds.) (1971). *Animals, Men and Morals: An Inquiry into the Maltreatment of Non-Humans*, Grove Press, New York.
- Goodin, R. (1992). *Green Political Theory*, Polity Press, London.
- Gorr, M. (1990). Thomson and the Trolley Problem. *Philosophical Studies*, 59(1):91–100.
- Grau, C. (2010). Moral Status, Speciesism and Liao’s Genetic Account. *Journal of Moral Philosophy*, 7(3).
- Greene, J. D. (2002). *The Terrible, Horrible, No Good, Very Bad Truth About Morality and What To Do About It*. Dissertation. Department of Philosophy, Princeton University.
- Greene, J. D. (2008). The Secret Joke of Kant’s Soul. In: Sinnott-Armstrong, W. (ed.) *Moral Psychology, Vol. 3: The Neuroscience of Morality*. Cambridge, MA: MIT Press.
- Greene, J., Nystrom L., Engell A., Darley J. & Cohen J. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*, 44:389–400.
- Greene, J., Sommerville R., Nystrom L., Darley J., & Cohen J. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, 293:2105–2108.
- Greenwald, A.G., & Banaji, M.R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, 102:4–27.
- Greenwald, A.G., McGhee, D.E., & Schwartz, J.K.L. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, 74:1464–1480.
- Griffin, D. (2001). *Animal Minds: Beyond Cognition to Consiousness*, University of Chicago Press.

- Haack, S. (1993). *Evidence and Inquiry*, Oxford, UK: Blackwell.
- Haidt, J. (2001). The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment. *Psychological Review*, 108(4): 814-834.
- Haidt, J. (2012). *The Righteous Mind*. Pantheon.
- Hammond, P. (1976). Why ethical measures of inequality need interpersonal comparisons. *Theory and Decision*, 7:263-274.
- Hanna, R. (1992). Morality De Re: Reflections on the Trolley Problem. In: Fischer, J. M. & Ravizza, M., (eds.) *Ethics: Problems and Principles*. New York: Holt, Rinehart & Winston.
- Hare, R.M. (1991). *The Language of Morals*, Clarendon.
- Hargreaves-Heap, S.H., Varoufakis, Y. (2004). *Game Theory: a Critical Text*. London and New York: Routledge.
- Harris, J. (2000). The moral difference between throwing a trolley at a person and throwing a person at a trolley: a reply to Kamm. *Proceedings of the Aristotelian Society, Supplementary Volumes*, Vol. 74 pp.41-57.
- Harris, S. (2004). *The End of Faith*. W.W.Norton.
- Harris, S. (2010). *The Moral Landscape: How Science Can Determine Human Values*. Free Press.
- Harsanyi, J.C. (1953): Cardinal Utility in Welfare Economics and in the Theory of Risk-Taking. *Journal of Political Economy*, 61:434-435.
- Harsanyi, J.C. (1955). Cardinal welfare, individualistic ethics, and the interpersonal comparison of utility. *Journal of Political Economy*, 63:309-321.
- Hauser, M., Young L., & Cushman F. (2008). 'Reviving Rawls' Linguistic Analogy: Operative Principles and the Causal Structure of Moral Actions. In: Sinnott-Armstrong W. (ed.) (2008), *Moral psychology and biology*. Oxford University Press, New York.
- Herzog, H. (2010), *Some We Love, Some We Hate, Some We Eat: Why It's So Hard to Think Straight About Animals*, Harper.
- Hoekstra, A.Y. (2010). The water footprint of animal products, In: D'Silva, J. and Webster, J. (eds.) *The meat crisis: Developing more sustainable production and consumption*, Earthscan, London, UK, pp.22-33.
- Holtug, N. (2006). Prioritarianism. In: *Egalitarianism, new essays on the nature and value of equality*, Holtug & Lippert-Rasmussen (eds.), Clarendon Press.
- Holtug, N. (2007). Equality for Animals. In J. Ryberg, T.S. Petersen & C. Wolf (eds.), *New Waves in Applied Ethics*. Palgrave Macmillan.
- Hooker, B. (2011). Rule Consequentialism, *The Stanford Encyclopedia of Philosophy* (Spring 2011 Edition), Edward N. Zalta (ed.).
- Horta, O. (2010). The Ethics of the Ecology of Fear against the Nonspeciesist Paradigm: A Shift in the Aims of Intervention in Nature. *Between the species*, Issue X:163-187.
- Horta, O. (2010b). What is Speciesism? *The Journal of Agricultural and Environmental Ethics*, 23:243-266.
- Horta, O. (2010c). Debunking the idyllic view of natural processes: population dynamics and suffering in the wild. *Télos*, 17:73-88.
- Howard-Snyder F. (2011). Doing vs. Allowing Harm. *The Stanford Encyclopedia of Philosophy* (Winter 2011 Edition), Edward N. Zalta (ed.).
- Hull, D. (1986). On Human Nature. *PSA: Proceedings of the biennial meeting of the philosophy of science association*, 2: 3-13.
- Hurka, T. (1983). Value and Population Size. *Ethics*, 93: 496-507.
- Hursthouse, R. (2000). *Ethics, Humans and Other Animals*, Routledge, London.
- Huther, C. (2005). *Can Speciesism be Defended? A Discussion of the Traditional Approach to the Moral Status of Animals*. Munich: Ludwig-Maximilians-Universität München.
- Jamieson, D. (1990). Rights, Justice and the Duty to Provide Assistance: A Critique of Regan's Theory of Rights. *Ethics*, 100(2): 349-62.
- Jensen, K.K. (2003). What is the Difference Between (Moderate) Egalitarianism and Prioritarianism? *Economics and Philosophy*, 19 (1), pp.89-109.

- Joy, M. (2002). *Psychic Numbing and Meat Consumption: The Psychology of Carnism*, Ph.D. dissertation, Saybrook Graduate School.
- Joy, M. (2009). *Why we love dogs, eat pigs and wear cows, An introduction to carnism*, Conari Press.
- Kagan, S. (1999). *Equality and Desert*. In: L.P. Pojman and O. McLeod (eds.), *What Do We Deserve? A Reader on Justice and Desert*, Oxford and New York: Oxford University Press, pp. 298-314.
- Kahneman, D. (2003). Maps of Bounded Rationality: Psychology for Behavioral Economics. *American Economic Review*, 93 (5): 1449-1475.
- Kahneman, D. (2011). *Thinking, Fast and Slow*, Macmillan.
- Kahneman, D. & Shane, F. (2002). Representativeness Revisited: Attribute Substitution in Intuitive Judgment. In Thomas Gilovich, Dale Griffin, Daniel Kahneman. *Heuristics and Biases: The Psychology of Intuitive Judgment*. Cambridge: Cambridge University Press. pp. 49-81.
- Kahneman, D. & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, XLVII, 263-291.
- Kahneman, D. & Tversky, A. (1984). Choices, Values, and Frames. *American Psychologist*, 39 (4): 341-350.
- Kahneman, D., Tversky, A. & Slovic, P., eds. (1982). *Judgment under Uncertainty: Heuristics & Biases*. Cambridge, UK, Cambridge University Press.
- Kamm, F.M. (1989). Harming Some to Save Others. *Philosophical Studies*, 57:227-60.
- Kamm, F.M. (1998). *Morality, Mortality: Death and Whom to Save from It*. New York: Oxford University Press.
- Kamm, F.M. (2000). The Doctrine of Triple Effect and Why a Rational Agent Need Not Intend the Means to His End. *Proceedings of the Aristotelian Society*, Supplementary Volumes, 74:21-39.
- Kamm, F.M. (2007). *Intricate Ethics. Rights, Responsibilities, and Permissible Harm*, Oxford University Press. Oxford.
- Kant, I. (1785), translated by J.W. Ellington (1993). *Grounding for the Metaphysics of Morals 3rd ed.* Hackett.
- Karban, R. & K. Shiojiri (2009). Self-recognition affects plant communication and defense. *Ecology Letters*, 12.
- Kaufman, F. (1998). Speciesism and the Argument from Misfortune, *Journal of Applied Philosophy*, 15 (2): 155-163.
- Kerstein S. (2009). Treating Others Merely as Means. *Utilitas*, 21(2):163-180.
- Kolmogorov, A. (1930). On the Notion of Mean. In: *Mathematics and Mechanics*. Kluwer, pp. 144-146.
- Korsgaard, C. (1996). The Right to Lie: Kant on Dealing with Evil. In *Creating the Kingdom of Ends*. Cambridge: Cambridge University Press.
- Kuhn, T. (1962). *The Structure of Scientific Revolutions*. The University of Chicago Press.
- Kuhn, T. (1970). Objectivity, Value Judgment, and Theory Choice. In: *Science, Reason, and Reality*. New York: Harcourt Brace, 1998.
- Kurzban, R., Tooby, J. & Cosmides, L. (2001). Can race be erased? Coalitional computation and social categorization. *PNAS*, 98(26):15387-15393.
- Lamey, A. (2007). Food fight! Davis versus Regan on the ethics of eating beef. *Journal of Social Philosophy*, Vol. 38 No. 2,
- Lamont, J. (1994). The Concept of Desert in Distributive Justice. *The Philosophical Quarterly*, 44: 45-64.
- Lanteri, A., Chelini, C., & Rizzello S. (2008). An Experimental Investigation of Emotions and Reasoning in the Trolley Problem. *Journal of Business Ethics*, 83:789-804.
- Lee, P. & George, R. (2008). The Nature and Basis of Human Dignity. *Ratio Juris*, 21(2):173-93.

- Liao, M. (2009). The Loop Case and Kamm's Doctrine of Triple Effect. *Philosophical Studies*, 146(2):223-231.
- Liao, M. (2010). The Basis of Human Moral Status. *Journal of Moral Philosophy*, 7(2): 159-179.
- Liao, M., Wiegmann, A., Alexander, J. & Vong, G. (2011). Putting the Trolley in Order: Experimental Philosophy and the Loop Case. *Philosophical Psychology*, 25(5):1-11.
- Lippert-Rasmussen, K. (1996). Moral Status and the Impermissibility of Minimizing Violations. *Philosophy & Public Affairs*, 25(4), pp. 333-351.
- Loomis, E.S. (1968). *The Pythagorean Proposition* (2nd ed.). The National Council of Teachers of Mathematics.
- Lumer, C. (2006). Prioritarian Welfare Functions. An Elaboration and Justification. In: Daniel Schoch (ed.): *Democracy and Welfare*. Paderborn: Mentis.
- Margolis, H. (1987). *Patterns, Thinking and Cognition: A Theory of Judgement*. Chicago: University of Chicago Press.
- Maslow, A., (1943). A Theory of Human Motivation. *Psychological Review*, Vol. 50, nr.4, pp. 370-396.
- Masson, J.M. (1995). *When Elephants Weep: The Emotional Life of Animals*, Jeffrey M. Masson, Delta.
- Matheny, G. (2003). Least harm: A defense of vegetarianism from Steven Davis's omnivorous proposal. *Journal of Agricultural and Environmental Ethics*, 16: 505-511.
- Matheny, G. & Chan, K.M.A. (2005). Human Diets and Animal Welfare: The Illogic of the Larder. *Journal of Agricultural and Environmental Ethics*, 18: 579-594.
- McCarthy, D. (2003). *Prospects for Prioritarianism*. Unpublished
- McCarthy, D. (2006). Utilitarianism and Prioritarianism I. *Economics and Philosophy*, 22:335-63.
- McCarthy, D. (2008). Utilitarianism and Prioritarianism II. *Economics and Philosophy*, 24:1-33.
- McCloskey, H. J. (1965). A Non-Utilitarian Approach to Punishment. *Inquiry*, 8:239-255.
- McDowell, J. (1984). Values and Secondary Qualities. In Honderich, T. (ed.), *Morality and Objectivity*. Routledge.
- McIntyre, A. (2001). Doing Away with Double Effect, *Ethics*, Vol. 111, No. 2, pp. 219-255.
- McIntyre A. (2011). Doctrine of Double Effect. *The Stanford Encyclopedia of Philosophy (Fall 2011 Edition)*, Edward N. Zalta (ed.)
- McMahan, J. (2002). *The Ethics of Killing: Problems at the Margins of Life*. Oxford: Oxford University Press.
- McMahan, J. (2005). Our Fellow Creatures. *The Journal of Ethics*, 9: 353-380.
- McMahan, J. (2008). Eating Animals the Nice Way. *Daedalus*, Winter 2008.
- McMahan, J. (2009). Asymmetries in the Morality of Causing People to Exist. In Roberts M. & Wasserman, D. (eds), *Harming Future Persons*. International Library of Ethics, Law, and the New Medicine, Volume 35.
- Mikhail, J. (2000). *Rawls' linguistic analogy: A Study of the 'generative grammar' model of moral theory described by John Rawls in 'A theory of justice.'* PhD dissertation, Cornell University, Ithaca, NY.
- Mikhail, J. (2007). Universal Moral Grammar: Theory, Evidence and the Future. *Trends in Cognitive Sciences*, 11(4): 143-152.
- Miller, D. (1976). *Social Justice*, Oxford: Clarendon Press.
- Milne, H. (1986). Desert, effort and equality. *Journal of Applied Philosophy*, 3: 235-243.
- Montmarquet, J. (1982). Doing Good: The Right and the Wrong Way. *Journal of Philosophy*, LXXIX:439-455.
- Moreno-Ternero, J.D. & Roemer, J.E. (2005). *Impartiality and Priority. Part 1: the Veil of Ignorance*. Working paper.
- Mulgan, T. (2006). *Future People. A Moderate Consequentialist Account of our Obligations to Future Generations*, Oxford: Clarendon Press.
- Müller-Lyer, F.C. (1889). Optische Urteilstäuschungen. *Archiv für Physiologie Suppl.*, 263-270.
- Narveson, J. (1967). Utilitarianism and New Generations. *Mind*, 76: 62-72.

- Nelkin, D.K. (2013), Moral Luck, *The Stanford Encyclopedia of Philosophy* (Summer 2013 Edition), Edward N. Zalta (ed.).
- Ng, Y-K. (1986). Social Criteria for Evaluating Population Change: an Alternative to the Blackorby-Donaldson Criterion. *Journal of Public Economics*, 29:375-381.
- Nobis, N. (2004). Carl Cohen's 'Kind' Arguments For Animal Rights and Against Human Rights. *Journal of Applied Philosophy*, 21 (1):43-59.
- Noddings, N. (2002). *Starting at Home: Caring and Social Policy*, Berkeley, CA.: University of California Press.
- Nolt, J. (2013). Comparing Suffering Across Species. *Between the Species*, 16(1):86-104.
- Norcross, A. (2008). Off Her Trolley? Frances Kamm and the Metaphysics of Morality. *Utilitas*, 20(1): 65-80.
- Nozick, R. (1974). *Anarchy, State, and Utopia*. Basic Books, New York.
- Nussbaum, M. (1992). Human Functioning and Social Justice: In Defense of Aristotelian Essentialism. *Political Theory*, 20:202-246.
- Nussbaum, M. (2000). *Women and Human Development: The Capabilities Approach*, Cambridge: Cambridge University Press.
- Nussbaum, M. (2006) *Frontiers of Justice. Disability, Nationality, Species Membership*, The Belknap Press of Harvard University, Cambridge, Mass.
- Olewski, J. (2010). Calculating our Nitrogen Footprint. Efficiency of nitrogen use in different farming systems on a global scale, *Growing Green International*, Nr.26.
- O'Neill, P., & Petrinovich, L. (1998). A Preliminary Cross-cultural Study of Moral Intuitions. *Evolution and Human Behavior*, 19(6): 349-367.
- Oswald, J. (1791). *The Cry of Nature; or, An Appeal to Mercy and to Justice, on Behalf of the Persecuted Animals*.
- Otsuka, M. (2008). Double Effect, Triple Effect and the Trolley Problem: Squaring the Circle in Looping Cases. *Utilitas*, 20(1):92-110.
- Otsuka, M. & Voorhoeve, A. (2009). Why It Matters That Some Are Worse Off Than Others: An Argument against the Priority View. *Philosophy & Public Affairs*, 37(2):171-99.
- Parfit, D. (1984). *Reasons and Persons*, Oxford: Clarendon Press.
- Parfit, D. (1991). *Equality or Priority, The Lindlev Lecture*, Lawrence: University of Kansas.
- Parfit, D. (1997). Equality and Priority. *Ratio*, 10:202-221.
- Parfit, D. (2011). *On What Matters*. Oxford University Press.
- Parks, B. D. (2006). The Natural-Artificial Distinction and Conjoined Twins: A Response To Judith Thomson's Argument for Abortion Rights. *National Catholic Bioethics Quarterly*, 6(4):671-680.
- Petrinovich, L., & O'Neill, P. (1996). Influence of Wording and Framing Effects on Moral Intuitions. *Ethology and Sociobiology* 17:145-171.
- Phillips C., Wagers W. and Lau, E.F. (2010). Grammatical Illusions and Selective Fallibility in Real-Time Language Comprehension. In J. Runner (ed.), *Experiments at the Interfaces, Syntax & Semantics*, vol. 37. Emerald Publications.
- Pivato, M. (2009). *Social Choice with Approximate Interpersonal Comparisons of Well-being*. MPRA Paper No. 17222.
- Plous, S. (2003). Is there such a thing as prejudice towards animals? In *Understanding prejudice and discrimination*, Plous (ed.), McGraw-Hill, New York.
- Plumwood, V. (1993). *Feminism and the Mastery of Nature*, Routledge, London.
- Pohl, R. (2004). *Cognitive Illusions: A Handbook on Fallacies and Biases in Thinking, Judgement and Memory*. Psychology Press.
- Porter, T. (2012). In Defense of the Priority View. *Utilitas*, 24:349-365.
- Postow, B.C. (1989). Thomson and the Trolley Problem. *Southern Journal of Philosophy*, 27(4):529-537.
- Pratt, J. W. (1964). Risk aversion in the small and in the large. *Econometrica* 32:122-136.

- Purves, D. & Lotto, B. (2002). *Why We See What We Do: An Empirical Theory of Vision*, Sunderland, MA: Sinauer Associates.
- Pylyshyn, Z. (1999). Is vision continuous with cognition? The case for cognitive impenetrability of visual perception. *Behav Brain Sci.*, 22:341-65.
- Quattrone, G.A. & Jones, E.E. (1980). The perception of variability within in-groups and out-groups: Implications for the law of small numbers. *Journal of Personality and Social Psychology*, 38(1):141-152.
- Quinn, W. (1989a) Actions, Intentions, and Consequences: The Doctrine of Doing and Allowing. *Philosophical Review* 98:287-312.
- Quinn, W. (1989b) Actions, Intentions, and Consequences: The Doctrine of Double Effect. *Philosophy and Public Affairs*, 18(4):334-351.
- Rabinowicz, W. (2002). Prioritarianism for Prospects. *Utilitas*, 14:2-21.
- Rachels, J. (1990). *Created From Animals. The Moral Implications of Darwinism*. Oxford University Press.
- Rai, T. & Holyoak, K. (2010). Moral Principles or Consumer Preferences? Alternative Framings of the Trolley Problem. *Cognitive Science* 34:311-321.
- Ramachandran, V.S. and Hubbard, E.M. (2001). Synaesthesia — a window into perception, thought and language. *Journal of Consciousness Studies*, 8: 3-34.
- Rawls, J. (1971). *A Theory of Justice*. Cambridge: Harvard University Press.
- Rawls, J. (2001). *Justice as Fairness: A Restatement*, ed. by E. Kelly, Cambridge: Harvard University Press.
- Reibetanz, S. (1998). A Problem for the Doctrine of Double Effect. *Proceedings of the Aristotelian Society*, New Series, Vol. 98, pp. 217-223
- Regan, T. (1983). *The Case for Animal Rights*, Berkeley: University of California Press.
- Riley, J. (1989). Justice Under Capitalism. In: *Markets and Justice*, ed. J.W. Chapman, New York: New York University Press, 122-162.
- Ritson, J. (1802). *An Essay on Abstinence from Animal Food as a Moral Duty*.
- Roemer, J.E. (1996). *Theories of Distributive Justice*, Cambridge: Harvard University Press.
- Rolston III, H. (1988). *Environmental Ethics: Duties to and Values in the Natural World*. Philadelphia: Temple University Press.
- Rosenberg, M. (2003). *Nonviolent Communication: A Language of Life*. 2nd Edition. Encinitas, CA: PuddleDancer Press.
- Ross, W. D. (1930, 2002). *The Right and the Good*. (With an introduction and bibliography by Philip Stratton-Lake). Oxford University Press.
- Rowlands, M. (1997). Contractarianism and Animal Rights. *Journal of Applied Philosophy*, 14(3):235-247.
- Rowlands, M. (1998). *Animal Rights: A Philosophical Defence*, Macmillan/St Martin's Press.
- Rubin, M. & Badea, C. (2012). They're all the same!...but for several different reasons: A review of the multicausal nature of perceived group variability. *Current Directions in Psychological Science*, 21: 367-372.
- Russell, B. (1979). The Relative Strictness of Positive and Negative Duties. In: Steinbock, B. (ed.) *Killing and Letting Die*. NY: Prentice-Hall.
- Ryder, R.D. (1975). *Victims of Science: The Use of Animals in Research*, Davis-Poynter.
- Ryder, R.D. (2001). *Painism: A Modern Morality*, Opengate Press.
- Sachdeva, S. e.a. (2011). Culture and the quest for universal principles in moral reasoning. *International Journal of Psychology*, 46(3).
- Sadurski, W. (1985). *Giving Desert Its Due*, Dordrecht, Holland: D. Reidel.
- Salt, H. (1892). *Animals' Rights: Considered in Relation to Social Progress*.
- Sapontzis, S. (1984). Predation. *Ethics and animals*, 5(2):27-38.
- Scanlon, T.M. (1998). *What We Owe to Each Other*. Cambridge, Mass.; Belknap Press.
- Scanlon, T.M. (2008). *Moral Dimensions*. Harvard University Press.

- Scheffler, S. (1982). *The Rejection of Consequentialism*, Oxford and New York: Oxford University Press.
- Schnall, S., Haidt, J., Clore, G.L., Jordan, A.H. (2008). Disgust as Embodied Moral Judgment. *PSPB*, 34(8): 1096-1109.
- Schwitzgebel, E. & Cushman F. (2012). Expertise in Moral Reasoning? Order Effects on Moral Judgment in Professional Philosophers and Non-Philosophers. *Mind & Language* 27:135-153.
- Scruton, R. (1998). *Animal Rights and Wrongs*, Demos.
- Segall, M.H., Campbell, D.T. & Herskovits, M.J. (1963). Cultural Differences in the Perception of Geometric Illusions. *Science*, New Series, Vol. 139, No. 3556, pp. 769-771.
- Sen, A. (1982). *Choice, Welfare and Measurement*, MIT Press.
- Sen, A. (1992). *Inequality Reexamined*, Cambridge: Harvard University Press.
- Serpell, J. (1996). *In the Company of Animals, a study of human-animal relationships*, Cambridge University Press.
- Shaver, R. (2011). Thomson's Trolley Switch, *Journal of Ethics & Social Philosophy*, discussion note.
- Shaw, W.H. (1999). *Contemporary Ethics: Taking Account of Utilitarianism*. Oxford: Blackwell.
- Shaw, J. (2006). Intentions and Trolleys, *Philosophical Quarterly*, 56 (222):63 - 83.
- Shermer, M. (2004). *The Science of Good and Evil*. New York: Times Books.
- Shiell, L. (2005). *The Repugnant Conclusion on Realistic Choice Sets*, University of Ottawa.
- Simmons, A. (2009). Animals, predators, the right to life and the duty to save lives. *Ethics & the environment*, 14(1):15-27.
- Singer, P. (1973). *Food for Thought* [Reply to David Rosinger]. New York Review of Books 20.10 1973.
- Singer, P. (1975, 1990). *Animal Liberation, a new ethics for our treatment of animals*. 2nd ed., New York Review of Books.
- Singer, P. (1993). *Practical ethics*. Cambridge University Press.
- Singer, P. (2005). Ethics and Intuitions. *The Journal of Ethics* 9:331-352.
- Singer, P. & Cavalieri, P. (eds.) (1993). *The Great Ape Project: Equality Beyond Humanity*. Fourth Estate publishing, London, England.
- Sinnott-Armstrong, W. (2008). Framing Moral Intuitions. In W. Sinnott -Armstrong (Ed.) *Moral Psychology, Volume 2: The Cognitive Science of Morality*, (pp. 47-76). Cambridge, MA: MIT Press.
- Sinnott-Armstrong, W., Young, L., & Cushman, F. A. (2010). Moral Intuitions as Heuristics. In J. Doris et al. (Eds.), *The Oxford Handbook of Moral Psychology*. Oxford University Press.
- Slote, M. (2001). *Morals from Motives*. Oxford: Oxford University Press.
- Smart, J.C.C. & Williams, B. (1973). *Utilitarianism: For and Against*. Cambridge University Press.
- Sneddon, L.U., Braithwaite, V.A. and Gentle, M.J. (2003a). Do fishes have nociceptors? Evidence for the evolution of a vertebrate sensory system. *Proceedings Of The Royal Society Of London Series B Biological Sciences*, 270 (1520): 1115-1121.
- Sneddon, L.U., Braithwaite, V.A. and Gentle, M.J., (2003b). Novel object test: Examining nociception and fear in the rainbow trout. *Journal Of Pain*, 4 (8): 431-440.
- Sobel, D. (2007). The Impotence of the Demandingness Objection. *Philosophers' Imprint*, 7(8):1-17.
- Sunstein, C. (2005). Moral Heuristics. *Behavioral and Brain Sciences*, 28:531-573.
- Tajfel, H. (1981). *Human Groups and Social Categories*. Cambridge University Press, Cambridge.
- Tanner, J. (2006). Marginal humans, the argument from kinds and the similarity argument. *Facta Universitas*, 5(1):47-63.
- Tanner, J. (2009). The Argument From Marginal Cases and the Slippery Slope Objection. *Environmental Values*, 18:51-66.
- Taylor, P. (1986). *Respect for Nature: A Theory of Environmental Ethics*. Princeton University Press.

- Temkin, L. (1987). Intransitivity and the Mere Addition Paradox. *Philosophy and Public Affairs*, 16 (2): 138–187.
- Thomson, J.J. (1971). A Defense of Abortion. *Philosophy and Public Affairs*, 1(1):47–66.
- Thomson, J.J. (1976). Killing, Letting Die, and the Trolley Problem. *The Monist*, 59:204–217.
- Thomson, J.J. (1985). The Trolley Problem. *The Yale Law Journal*, 94:1395–415.
- Thomson, J.J. (1993). Reply to Commentators. *Philosophy and Phenomenological Research*, 53(1):187–194.
- Thomson, J.J. (2008). Turning the Trolley. *Philosophy and Public Affairs*, 36(4):359–374.
- Unger, P. (1996). *Living High and Letting Die*, Oxford: Oxford University Press.
- Valdesolo, P. & DeSteno, D. (2006). Manipulations of Emotional Context Shape Moral Judgment. *Psychological Science*, 17:476–77.
- Vallentyne, P. (2006). Of Mice and Men: Equality for Animals. In: *Egalitarianism, new essays on the nature and value of equality*, Holtug & Lippert-Rasmussen (eds.), Clarendon Press.
- Vallentyne, P. (2012). Libertarianism. *The Stanford Encyclopedia of Philosophy (Spring 2012 Edition)*, Edward N. Zalta (ed.)
- Van den Berg, F. (2011). *Harming Others: Universal Subjectivism and the Expanding Moral Circle*. PhD dissertation, Universiteit Leiden.
- Van De Veer, D. (1979). Of Beasts, Person, and the Original Position, *The Monist*, 62:368–377.
- Velleman, D. (1991). Well-being and Time. *Pacific Philosophical Quarterly*, 72:50.
- Visak, T. (2011). *Killing Happy Animals. Explorations in Utilitarian Ethics*. Dissertation, Utrecht University.
- Waldmann M. & Dieterich J. (2007). Throwing a bomb on a person versus throwing a person on a bomb. Intervention myopia in moral intuitions, *Psychological Science*, 18(3):247–253.
- Weinberg, S. (1996). *The Quantum Theory of Fields*. Cambridge University Press.
- Weirich, P. (1983). Utility Tempered with Equality. *Nous*, 17: 423–39.
- Weymark, J.A. (1981). Generalized Gini Inequality Indices. *Mathematical Social Sciences*, 1:409–430.
- Whitley, B.E., & Kite, M.E. (2010). *The Psychology of Prejudice and Discrimination*. Belmont, CA: Wadsworth.
- Williams, B. (1970). The Self and the Future, *The Philosophical Review*, 79(2):161–180.
- Wilson, S. (2001). Carruthers and the Argument from Marginal Cases. *Journal of Applied Philosophy*, 18 (2):135–147.

